Capstone Proposal

Customer Segmentation and future customer prediction with Arvato financial solutions

Proposal Date: Aug 17th, 2022
Author: Ahmed Khaled Metwally

Udacity Machine Learning Engineer Nanodegree

Project Overview

The project is divided into two main components. The first part is the customer segmentation and differentiating the customers demographics that are considered the core of the company. The second part is utilising the first part to be able to predict the customers that will most likely be customers for the company.

Machine learning will be used for the following underlying business matter: a client mail-order company trying to acquire more customers more efficiently.

Customer grouping will enable the company's marketing team to decide: for which demographic part will the promotional campaign be most appropriate.

Problem Statement

In this project, the problem is how the mail-order company acquires new customers more efficiently, given the German demographic data. This problem is divided into two tasks: customer segmentation using unsupervised learning techniques and the probability of being a new customer of the company using supervised learning.

- The unsupervised learning methods on the customers data and their demographics will result in creating customer segments.
- After that we can train a model to predict the probability of a person in becoming a new customer (given a certain threshold) using supervised learning methods on another dataset.

Training Data (Datasets)

There will be four datasets in this project (two for unsupervised and another two training and testing):

- 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).

Methodology

The project implementation will be on Sagemaker and data storage in S3.

From the current customer base, in order to better target the German population, it is required to predict in advance who would become a future customer.

Performing unsupervised learning will identify customer segments that can be done by first applying the "*Principle Component Analysis*" (PCA) for dimensionality reduction as the datasets provided have too many features. In the dataset, the rows represent observations required to be embedded in a lower dimensional space. The columns represent the features required for finding a reduced approximation for. The algorithm that will be used will find the covariance matrix, and then perform the singular value decomposition to produce the principal components. The "*K-Means*" Clustering will come into play next to group the observations with similar attribute values (the points corresponding to these observations are closer together). Where the *n* attributes (people in the azdias dataset for example) in each row represent a point in *n*-dimensional space (the principle components that result from the PCA). The Euclidean distance between these points represents the similarity of the corresponding observations and hence groups similar clusters.

Then, for the second part, supervised learning will be used which is responsible for predicting future potential clients from the German Population dataset. Different supervised algorithms can be of use such as XGBoost classifier or Random Forests.

We can then test this classification (either the person is a customer or not) using logistic regression as it is a binary classification problem.

Utilising the Area Under the Curve (AUC) for the ROC curve will define the ranking in the kaggle competition. According to the competition's evaluation section[1]: The

[1] https://www.kaggle.com/competitions/udacity-arvato-identify-customers/overview/evaluation

evaluation metric for this competition is AUC for the ROC curve[2], relative to the detection of customers from the mail campaign. A ROC, or receiver operating characteristic, is a graphic used to plot the true positive rate (TPR, proportion of actual customers that are labelled as so) against the false positive rate (FPR, proportion of non-customers labelled as customers).

Design

Saving datasets in "*S3*" and using "*Sagemaker*" to do the following:

1. Data Exploration and Cleaning
    a. Exploring the dataset and trying to detected biases and null values
    b. Cleaning the raw input datasets
2. Data Visualisation
    a. Develop the research questions more clearly
    b. Find correlations between the data
3. Feature Engineering
    a. This is the stage where PCA should be used to make informed decisions about features
4. Model Selection and Training
    a. Unsupervised learning can be done using K-means or non-linear fitting with kernels and whichever acts best will be chosen as a clustering algorithm
    b. Experimenting XGBoost and decision trees and decide which will be used
5. Model Prediction
    a. Testing the model prediction will act as the benchmark

[2] https://en.wikipedia.org/wiki/Receiver_operating_characteristic#Area_under_the_curve