Capstone project

Machine Learning Engineer Nanodegree

Ganesh Bhat
25 April 2021

Project Overview (Domain Background)

Customer Segmentation Report for Arvato Financial Services

In this project, we will analyze demographics data for customers of a mail-order sales company in Germany, comparing it against demographics information for the general population.

We have been given the data-set as outlined below and we will use a combination of unsupervised and supervised learning techniques to answer few questions for the customer as outlined in the problem statement.

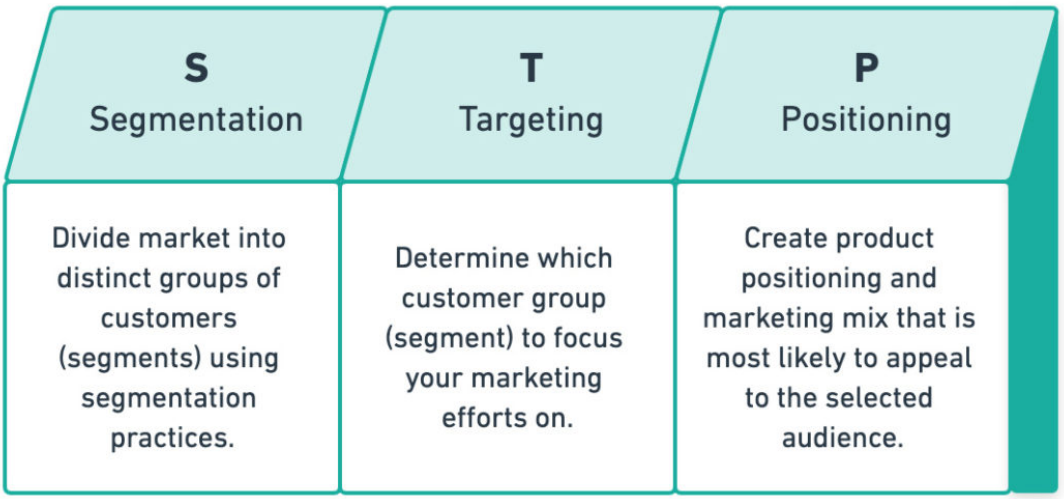

Image Reference

Personal motivation:

My personal curiosity stays with the time-series forecasting of financial data, personal finance, I wanted to apply the learnings from this course and do Machine Learning on Aws Sagemaker for this data.

This project nicely covers many areas such as PCA, Unsupervised and Supervised learning areas, Neural Networks and also provides an opportunity to apply all the different techniques such as Hyperparameter tuning jobs, balancing for certain metrics etc, which I learnt as part of this course.

I am looking at it as a practice project, where I can apply all my learnings and solidify my understanding of all the concepts, instead of focusing on one single larger problem at this stage.

Problem Statement

1. Identify the set of people mail-order sales company in Germany must target with their marketing campaign, instead of targeting all the people across germany.
2. Also identify customer segmentation, identify part of the population that best describes the core customer base (ex: regions, gender and income group etc) of people the company shall focus on, for their marketing campaign.
3. Company needs to know the prediction on which of those individuals are most likely to convert into becoming customers for the company so that they can target such individuals better using instruments like coupons, ads etc.

Datasets and Inputs

The dataset(s) and/or input(s) to be used in the project are thoroughly described. Information such as how the dataset or input is (was) obtained, and the characteristics of the dataset or input, should be included. It should be clear how the dataset(s) or input(s) will be used in the project and whether their use is appropriate given the context of the problem.

There were four data files provided by Arvato for this project.
(As part of the terms and conditions of Arvato, the files cannot be shared) This was obtained from Arvato, shared via Udacity for capstone project purpose.

Below describes the characteristic of the dataset.

- 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).

Each row of the demographics files represents a single person, but also includes information outside of individuals, including information about their household, building, and neighborhood.

The "CUSTOMERS" file contains three extra columns ('CUSTOMER_GROUP', 'ONLINE_PURCHASE', and 'PRODUCT_GROUP'), which provide broad information about the customers depicted in the file.

The original "MAILOUT" file included one additional column, "RESPONSE", which indicated whether or not each recipient became a customer of the company.

For the "TRAIN" subset, this column has been retained, but in the "TEST" subset it has been removed; it is against that withheld column that your final predictions will be assessed in the Kaggle competition.

Metrics

In the provided data the class is imbalanced, very similar to credit card fraud detection. We aim to use the AWS provided mechanism to add the balancing factor.

To evaluate unsupervised algorithm efficiency, we will use Silhouette score and measure the efficiency.

As we can see, adding additional individuals for ad-targeting will not have a major side effect. However leaving out an individual from ad-targeting can result in loss of customers.

Hence we will try to increase the ad-coverage than loosing them. We will tune our model for metric 'recall'.

An alternate possibility is to consider the metric such as ROC and AUR (reference)

Solution Statement

Solution is

1. Cluster Analysis: Perform dimensionality reduction to reduce the number of features using PCA and then apply a few unsupervised learning algorithms such as clustering (K-Means, DBScan) to show relationship between existing customers and rest of the people in germany. We will use elbow method to choose K for K-Means.

These could be clustered based on age,gender,income & occupation, cultural background, family status,

Various distance metrics such as Cosine Distance, Euclidean distance, Jaccard distance shall be considered.

2. Use one or many Supervised learning models to understand which model performs better in finding whether a person will become a customer or not - XGBoost, Neural Networks with sigmoid (Binary classification) .

Benchmark Model

We will use Logistic Regression or Naive Bayes for benchmarking purposes .
(Reference: Segmentation based modelling for advanced targeted marketing)

Reference: benchmarking predictive models

Evaluation Metrics

As we can see, adding additional individuals for ad-targeting will not have a major side effect. However leaving out an individual from ad-targeting can result in loss of customers. We will tune our model for metric 'recall'.

An alternate possibility is to consider the metric such as ROC and AUR (Reference: Beyond accuracy - precision and recall)

Project Design

Customer segmentation

- Load and prepare data
- Data exploration
- Data Cleaning
    - Handling missing values
    - Encoding categorical values
    - Feature Scaling
    - Remove highly co-related features

Unsupervised Learning & clustering

- Dimensionality reduction using PCA
- Analyze component makeup
- Apply unsupervised learning algorithm such as KMeans
- Choose correct K using elblow method (experiment with different k)
- Analyze the clusters using heatmap
- Use silhoutte score to evaluate the model

Supervised learning

- Load and prepare data
- Explore and preprocess data
- Split into train & test
- Prepare benchmark model and evaluate
- Use following models and check the accuracy
    - Neural Networks with PyTorch
    - XgBoost
    - KNN & SVM (optional, if time permits)
- Use hyperparameter tuner for betterment of results

- Evaluate the results using confusion metric and ROC score

- Visualize the predictions