Machine Learning Engineer Nanodegree

Arvato-Bertelsmann Customer Segmentation

Aakriti Sharma
September 30th, 2020

Domain Background

Maintaining healthy customer-relationships is of utmost importance for any company today. Therefore, it is crucial to strategically group customers into various groups based on demography and customer behavior. Customer segmentation helps us identify potential buyers, target the most profitable customers, and retain them over time. It can not only improve the marketing strategy but can also help in making better business decisions.

The success of any mailout marketing campaign depends on the customer list to which it is being targeted (Bult and Wansbeek, 1995; Bult et al., 1997; Dirk Van den Poel, 2003). Hence, it becomes critical to identify the target audience correctly and plan the strategies accordingly. The goal of this project will be to analyze the customer demographics and identifying the customer segments with the help of various machine learning algorithms.

Problem Statement

The goal of this project is to help a mail-order sales company in Germany to identify segments of the general population to target with their marketing to grow. The company has provided us with the demographic data of their current customers and the general population. We've to build a customer-segmentation machine learning model for the company, which correctly categorizes customers into groups and identifies the customers that the company should target.

Datasets and Inputs

There are four data files associated with 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).

The data required for the project has been provided by Bertelsmann Arvato Analytics.

Solution Statement

The solution to the above-defined problem can be broadly divided into two parts. First, comparing the general populating with the existing customers and identifying the segments of the general population that are more likely to be part of the company's main customer base, with the help of unsupervised learning techniques. Second is building a supervised learning model to predict the response of individual customers towards the marketing campaign. For the first part, we'll make use of demographics data of the general population and compare it to the customer demographic database, and for the later part, we'll use the demographic data with response labels of the customers to train the supervised learning model and make a prediction on our test set.

Benchmark Model

The benchmark performance for the given, over the Kaggle competition, problem statement has been about 0.80 AUC on the test set. To achieve this, various machine learning models, like K-means or Logistic Regression, can be applied and selected based on the defined metrics.

Evaluation Metrics

There will be mainly three evaluation metrics that will be used to evaluate the performance of the model:

1. **Silhouette score:** This metric will be used for the first part of the project to evaluate the customer segmentation model. It considers how well-separated groups are from each other. The silhouette coefficient *s* for a single sample is given as:

   $$s = \frac{b - a}{\max(a, b)}$$

   Where, **a**: the mean distance between sample and all other points in the same class,
   b: is the mean distance between a sample and all other points in the next nearest cluster.
   The higher value of silhouette represents higher cross-class separability and more tightly packed centroids.
2. **Accuracy:** For evaluating the supervised learning model accuracy will be used as an important metric. The formulation for which is given below:

   $$Accuracy = \frac{TP - TN}{TP + TN + FP + FN}$$

   Where, **TP:** True Positive, **TN:** True Negative, **FP:** False Positive, **FN:** False Negative.

3. **Area under the receiver operating curve (AUROC):** Since, there is a class imbalance, where most individuals don't respond to the mail out, thus AUROC metrics will be used for evaluating model performance on the training set as well as for the final test set.

An appropriate evaluation metric can only be chosen after the data has been analyzed properly.

Project Design

The proposed project design can be divided into the following steps:

1. **Data Exploration and Visualization:** Before building any model, it is important to understand our data by looking for missing values, noisiness in the data, type of distribution, and so on. In this step, I'd also visualize the data and study the correlation between the attributes and also identify the promising transformations that might be helpful.
2. **Data preprocessing:** After carefully analyzing the data, it is time to clean the data by filling in missing values, removing outliers, and so on, which is followed by feature engineering where we add new features or decompose the existing features and then normalize our data for the next step.
3. **Model training:** Here, the first step will be to build and compare various clustering algorithms for customer segmentation and to identify various groups in the population. After identifying segments, different classification algorithms like Logistic regression, decision trees, XGBoost can be used to predict if a person will become our next customer or not. We'll select the algorithms which performed better on the evaluation metrics.
4. **Fine-tuning model:** Now, we'll fine-tune our selected model using cross-validation. Model hyperparameters will be tuned using grid search.
5. **Model evaluation:** It is important to measure the performance of the model on the test and validation set, in order to avoid overfitting on the training set and build a more generalized model. After evaluation, we can finally make predictions on test data and submit them on Kaggle.

References

[1] Dirk Van den Poel, 2003, Predicting Mail-Order Repeat Buying: Which Variables Matter? Available at: http://ebc.ie.nthu.edu.tw/km/MI/crm/papper/wp_03_191.pdf

[2] Scikit-Learn Docs. Available at: https://scikit-learn.org/stable/modules/clustering.html#silhouette-coefficient