Create a Customer Segmentation Report for Arvato Financial Solutions

CAPSTONE PROPOSAL

Hussain M. Al-Balhareth

Udacity Machine Learning Nanodegree

Date: May-20-2021

Table of Contents

DOMAIN BACKGROUND

This project is connected to one of Udacity's capstone project options for the Machine Learning (ML) Engineer Nanodegree program, in connection with Arvato Financial Solutions, a Bertelsmann subsidiary.

In the project, a mail-order sales company in Germany is interested in identifying segments of the general population to target with their marketing in order to grow. Demographics information has been provided for both the general population at large as well as for prior customers of the mail-order company in order to build a model of the customer base of the company. The target dataset contains demographics information for targets of a mailout marketing campaign. The objective is to identify which individuals are most likely to respond to the campaign and become customers of the mail-order company.

As part of the project, half of the mailout data has been provided with included response column. For the competition, the remaining half of the mailout data has had its response column withheld; the competition will be scored based on the predictions on that half of the data. [1]

PROBLEM STATEMENT

In this project, I will analyze demographics data for customers of a mail-order sales company in Germany, comparing it against demographics information for the general population. I'll use unsupervised learning techniques to perform customer segmentation, identifying the parts of the population that best describe the core customer base of the company. Then, I'll apply what I've learned on a third 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. The data that I will use has been provided by Udacity partners at Bertelsmann Arvato Analytics, and represents a real-life data science task.

DATASETS AND INPUTS

There are four data files associated with this project:

1. Udacity_AZDIAS_052018.csv: Demographics data for the general population of Germany; 891 211 persons (rows) x 366 features (columns).
2. Udacity_CUSTOMERS_052018.csv: Demographics data for customers of a mail-order company; 191 652 persons (rows) x 369 features (columns). The three additional columns ['CUSTOMER_GROUP', 'ONLINE_PURCHASE', 'PRODUCT_GROUP'] provide broad information about the customers depicted in the file.
3. Udacity_MAILOUT_052018_TRAIN.csv: Demographics data for individuals who were targets of a marketing campaign; 42 982 persons (rows) x 367 (columns). The additional column ['RESPONSE'] indicates whether or not each recipient became a customer of the company.
4. Udacity_MAILOUT_052018_TEST.csv: Demographics data for individuals who were targets of a marketing campaign; 42 833 persons (rows) x 366 (columns). ['RESPONSE'] column has been removed so that the final predictions will be assessed in the Kaggle competition.

Furthermore, more information about the columns depicted in the files can be referred to in the two Excel spreadsheets below:

5. DIAS Information Levels - Attributes 2017.xlsx: a top-level list of attributes and descriptions, organized by informational category.
6. DIAS Attributes - Values 2017.xlsx: a detailed mapping of data values for each feature in alphabetical order.

SOLUTION STATEMENT

In order to solve this problem, I will go through three stages which are data cleaning and Exploratory Data Analysis (EDA), selecting and applying the optimum algorithms for both unsupervised and supervised machine learning models, and evaluating the final results against pre-defined metrics.

BENCHMARK MODEL

For unsupervised model, the over/under representation of customers data point within the general population clusters will be the indicator for the targeted audience. For the supervised models, the high scores within Kaggle Competition will be our benchmark.

EVALUATION METRICS

The PCA will be based on the amount of cumulative explained variance. The Elbow method will be used to select the optimum number of K-Means clusters. Besides training scores, cross-validation technique will be used to evaluate the overfitting of the supervised learning classifiers also utilizing ROC AUC scores.
As found in Part 2, there is a large output class imbalance, where most individuals did not respond to the mailout. Thus, predicting individual classes and using accuracy does not seem to be an appropriate performance evaluation method. Instead, the competition will be using AUC to evaluate performance. The exact values of the "RESPONSE" column do not matter as much: only that the higher values try to capture as many of the actual customers as possible, early in the ROC curve sweep.

PROJECT DESIGN

Data Analysis

To properly analyze the problem, I will perform an exploratory data analysis to better understand what algorithms and features are appropriate for solving it. As the datasets were not pre-cleaned, I will create a pre-processing function to clean four data files associated with this project, general population (AZDIAS), customers, and two MAILOUT files for targeted customers (TRAIN & TEST). The datasets contain information at individual and broader levels like household, building, neighborhood, etc.

Machine Learning Algorithms

Unsupervised ML Algorithm

Perform customer segmentation using the proper unsupervised learning methods such data transformation, dimensionality reduction using PCA, and K-means algorithm to analyze attributes

of established customers and the general population in order to create customer segments to be able to describe parts of the general population that are more likely to be part of the mail-order company's main customer base, and which parts of the general population are less so. The hope here is that certain clusters are over-represented in the customers data, as compared to the general population; those over-represented clusters will be assumed to be part of the core userbase. This information can then be used for further applications, such as targeting for a marketing campaign.

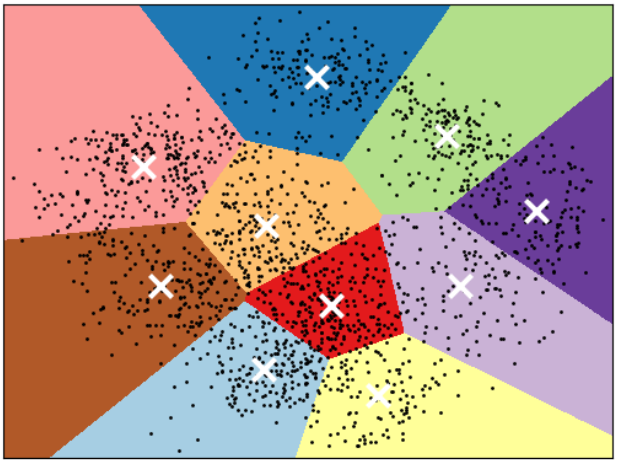

Figure: K-Means Clustering [2]

Supervised ML ***Model***

Access to a third MAILOUT dataset with attributes from targets of a mail order campaign. Use the previous analysis to build and evaluate a set of supervised machine learning models, such as Logistic Regression, RandomForestClassifier, DecisionTreeClassifier, or the popular XGBoost Classifier, and select the optimum one that best predicts whether or not each individual will respond to the campaign.

Results

Once I've chosen a model, I'll use it to make predictions on the campaign data as part of a Kaggle Competition link here. I'll rank the individuals by how likely they are to convert to being a customer, and see how my modeling skills measure up against my fellow students.

Blog Post

Finally, I will make a blog post to document all of the steps and decisions from start to finish of this project reporting all the findings.

REFERENCES

[1]. Kaggle.com. 2021. *Udacity+Arvato: Identify Customer Segments | Kaggle*. [online] Available at: <https://www.kaggle.com/c/udacity-arvato-identify-customers> [Accessed 20 May 2021].

[2]. Scikit-learn.org. 2021. *A demo of K-Means clustering on the handwritten digits data — scikit-learn 0.24.2 documentation*. [online] Available at: <https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_digits.html#sphx-glr-auto-examples-cluster-plot-kmeans-digits-py> [Accessed 20 May 2021].