Machine Learning Engineer Nanodegree

Capstone Proposal

Sohaib Zafar
April 28, 2020

Domain Background

In partnership with Bertelsmann Arvato, this project is part of the Udacity Machine Learning Engineer Nanodegree Capstone Project. For Arvato Financial Solutions, the main goal of this project is to build a Customer Segmentation Report.

Arvato offered demographic data on the general population of Germany as well as current mail-order customers (such as age, wages, wealth, education, properties, vehicles, homes, families, and so on). Under Arvato's terms and conditions, the data is secure and not accessible to the general public.

This demographic data will be used to classify mail-order company consumer segments in order to optimize targeted marketing efforts and forecast new customer conversion.

Problem Statement

This challenge is a real-life problem, provided by Arvato Financial Solutions, where the problem statement is:

How can their client, a mail-order company, gets new clients efficiently using data-driven approach for targeted marketing?

Customer Segmentation is performed using ML unsupervised learning techniques to classify the clusters of the population that best represent the company's core customer base.

After assessing the core customer base, marketing campaign data was used to apply ML supervised learning techniques to predict people who are likely to become new customers.

Data Sets

There are four data files associated with this project provided by Arvato

- 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).
- DIAS_Attributes_Values_2017.xlsx: Values- level information about attributes used in data.
- DIAS_Information_Levels_Attributes_2017_Komplett.xlsx: Top-level information about attributes used in data.

There are a number of missing values in these datasets, and not all of the features in a given Excel spreadsheet have a description, which needs to be discussed. Arvato presented a top-level list of attributes and descriptions, as well as a thorough mapping of data values for each element, in two spreadsheets.

Evaluation Metrics

Since this is a multi-class classification issue, the main metric to measure model output is the Area Under the Curve Receiver Operating Characteristics (ROC-AUC). The curve represents a measure of separability, with a higher score indicating better model efficiency. ROC-AUC protects against class imbalance, which is important in this case. The number of positive responses to an ad campaign is usually much lower than the number of negative responders. This is also the Kaggle submission's needed evaluation metric.

Benchmark

Before doing any hyper-parameter optimization, I will evaluate different baseline ML models and take cross validation score of Logistic Regression Model as a benchmark, before final model selection and parameter tuning.

Project Design

Part 0: Get to Know the Data

- Import the data
- Exploratory Data Analysis (EDA)
- Data Cleaning and Preprocessing -- Covert unknown values to NaN -- Convert special characters to NaN -- Delete columns with higher percentage of missing values
- Feature Engineering
- Feature Scaling (MinMax)

Part 1: Customer Segmentation Report

- Dimensionality Reduction using PCA
- K-means Clustering
- Customer Segmentation report using Cluster Mapping
- Component Makeup Analysis

Part 2: Supervised Learning Model

- Import the training data
- Perform data pre-processing steps: data cleaning, features engineering.
- Normalize the train data and save the scaler to be used later for test data.
- Select features based on feature importance
- Creating Baseline Models
- Selecting Model with high performance (using ROC-AUC metric) and low resource requirements
- Hyper-Parameter Tuning using Bayes Optimizer Cross Validation
- Save the model to be used in next step

Part 3: Kaggle Competition

- Import the testing data
- Perform data pre-processing steps: data cleaning, features engineering and scaling.
- Normalize the test data using scaler created for train data.-
- Make predictions using the model created in the supervised learning model part.
- Submit predictions to kaggle.