AWS Machine Learning Engineer Nanodegree

Capstone Proposal

Customer Segmentation for Arvato Financial Solutions

Domain Background

Arvato is a German services company that provides consulting services in the domain of financial services, information technology (IT), big data analytics, etc. It leveredges predictive analytics, leading-edge platforms and big data to deliver value to clients and help them see the bigger picture using data. It has a team of around 7000 IT, analytics and legal experts in 15 countries with a focus on Europe[1].

In this project, Arvato aims to utilize demographic data to identify the attributes of customers who convert and use this information to better target marketing outreach campaigns for a client (a mail-order company selling organic products) on people who have a higher propensity to covert and become a customer.

Problem Statement

The problem statement for this project can be summarized as given the data for customers and the general population, identify characteristics that differentiate customers. Also, use the customer information (characteristics like demographics, etc.) to predict whether the customer is likely to convert to a marketing campaign or not.

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).

Additional metadata is also provided to better understand the columns and what they represent.

Solution Statement

The solution to this project can be broken down into two major sub-parts -

1. Customer Segmentation - Using unsupervised techniques such as clustering segment the customer and the general population data to identify common characteristics of people who become customers, that is, respond positively to our campaigns.

2. Supervised Learning Model - Build a machine learning model to predict whether an individual (given his/her characteristics) is likely to convert to a customer using historical mail-order campaigns data.

Benchmark Model

We can use a logistic regression model with default model parameters as a benchmark model for our classification task (1 - customer conversion, 0 - non conversion). Using a simple logistic regression model is a common benchmark for classification tasks in machine learning.

Evaluation Metrics

We can use F1-Score[2] as an evaluation metric. This is because the dataset is highly imbalanced and using a standard metric like accuracy can give misleading results.

F1-Score takes into account both precision and recall and is defined as ->

$$F1\ Score\ =\ 2 * \frac{Precision * Recall}{Precision\ +\ Recall}$$

$$Precision\ =\ \frac{TP}{TP + FP},\ Recall\ =\ \frac{TP}{TP + FN}$$

$$Where,$$
$$TP\ =\ True\ Positive$$
$$FP\ =\ False\ Positive$$
$$FN\ =\ False\ Negative$$

Project Design

The following is a step-by-step outline of the proposed approach to complete this project -

1. Data Cleaning and EDA
    - Understand data and the fields available.
    - Clean the data to remove columns with large proportion of missing values, remove unrequired columns and impute any missing data.
    - Perform exploratory analysis on the data and get insights about the data distribution in terms of class distribution and other features.
2. Feature Engineering
    - Encode categorical features.
    - Standardize the data using standard scaler.
    - Perform PCA (Principal Component Analysis) to reduce the dimensionality in the data.
3. Customer Segmentation
    - Perform customer segmentation using a clustering algorithm like K-means clustering.
    - Determine the optimal number of clusters to use (using elbow method).
    - Visualize the distribution of the population and the customers across each cluster to identify clusters with higher customer concentration.
4. Model to predict customer conversion
    - Train a classification model to predict whether a customer would convert to a mail-order campaign using historical data.
    - Try different models like DecisionTreeClassifier, RandomForestClassifier, etc. and compare their performance to the baseline model
    - Use the trained model to make predictions on the test set.

References

1. Arvato Financial Services Website - https://finance.arvato.com/en/about-us/
2. F1 Score - Wikipedia - https://en.wikipedia.org/wiki/F-score