CAPSTONE PROPOSAL

The Bertelsmann Arvato mail order project

KIBRU TEMESGEN

MACHINE LEARNING ENGINEER NANODEGREE

DOMAIN BACKGROUND

Arvato is a service company that is wholly owned by Bertelsmann. It is one of eight divisions of Bertelsmann. Arvato develops and implements innovative solutions for business customers around the world. These include SCM solutions, financial services, and IT services, which are continuously developed with a focus on innovations in automation and data/analytics.

This is the same project on which the data scientist at Arvato is working. We have the attributes and the demographic information of the existing clients; we would analyze the attributes of the existing clients and match them against a bigger data set that includes the attributes for the people in Germany, and essentially figure out which people in Germany are more inclined to be their next customer for the clients of their mail-order service.

This real-life project will increase the efficiency of acquiring new customers. Instead of the mail-order company reaching out to all people in Germany, this will enable them to do a targeted campaign, which will reduce cost.

Problem Statement

The underlying problem to solve, as stated by the company representative, is ‘how can the mail-order company acquire new clients more efficiently?"

Given the behaviour and large amount of attributes of the existing customers, how can we bring more new customers from the general population is what we will answer in this

project. We will then predict the probability of a given person in germany joining as a new customer to our mail-ordering service.

Proposed Solution

The project will encompass two major parts. Firstly, segmenting the customer to understand the different demographics information and attributes of our existing customers and the general population. Secondly, making use of supervised learning techniques to train a classifier model on our existing customer. This model will be evaluated on our test data, reserved from our existing customer, and then used to make prediction for the general population of germany so that we know people who are likely to reply to our marketing.

Data sets and Inputs

Datasets and Inputs The project makes use of four files:

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).

3) Udacity_MAILOUT_052018_TRAIN.csv: Demographics data for individuals who were targets of a marketing campaign; 42 982 persons (rows) x 367 (columns).

4) Udacity_MAILOUT_052018_TEST.csv: Demographics data for individuals who were targets of a marketing campaign; 42 833 persons (rows) x 366 (columns).

Solution Statement

Arvato Financial Solutions' objective is to more effectively recruit new consumers in the German population by accurately forecasting who will become a customer from demographic profiles of current customers.

The first step to data cleaning is removing unwanted observations from the dataset. This includes duplicates, irrelevant data, and data that is not useful for the model or clustering. The next bucket under data cleaning involves fixing structural errors. This includes correcting typos, misspellings, and other errors that can affect the quality of the data. handling missing data by either removing the rows or columns that contain missing data or by imputing the missing values will also be a task to be done.

Since our data contains more than 350 different features, it is critical that we extract essential features. Wrapper method or principal component analysis will be the method used to minimize the number of features (PCA). After setting up the important feature attributes, we will segment clients using the unsupervised clustering algorithm k-means.

Proper supervised learning algorithm will be used to train a model and make prediction for the general population.

Benchmark Model

For the binary classification problem, a benchmark model would be a logistic regression model. To pick one supervised algorithm over another, the performance of this model may be compared to that of other algorithms.

Evaluation Metrics

The evaluation metrics used in the Bertelsmann Arvato mail order project were the AUC ROC curve. This value is used for the exact differentiation between the labels for the prediction. The AUC ROC curve is a graph showing the performance of a classification model at all classification thresholds. AUC stands for “Area under the ROC Curve.” That is, AUC measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1)1. The higher the AUC, the better the model’s performance at distinguishing between the positive and negative classes. . A ROC curve is a plot of the true positive rate (Sensitivity) in function of the false positive rate (100-Specificity) for different cut-off points of a parameter2. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold.

Project Design

First stage is data exploration, It involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors.

Next step is exploratory data analysis (EDA). This is a process where welook at and understand thedata with statistical and visualization methods. This step helps identify patterns and problems in the dataset, as well as deciding which model or algorithm to use in subsequent steps. Visualization tools will also be used to make the data more understandable.

Selecting the important feature to perform customer segmentation and supervised learning. Eventually, the model would be tested.

The task would be performed on aws sagemaker would be used for unsupervised/supervised learning.