Machine Learning Engineer Nanodegree

Capstone Proposal

John Anisere
January 1st, 2020

Proposal

Domain Background

The project is using the data provided by Arvato Financial Solutions, a Bertelsman subsidiary company, to perform customer segmentation and uncover the demographic characteristics of the core customers for the client of the company. We will also build a model by using another dataset from a marketing campaign of the company, to predict individuals that are most likely to be the potential customer of the company.

Problem Statement

I. Data Preprocessing: In this part we need to preprocess data for further analysis.Missing values by columns and rows will be analysed, data will be divided by types followed by subsequent transformations.

II. Customer Segmentation: In this part we need to analyze general population and customer segment data sets and use unsupervised learning techniques to perform customer segmentation, identifying the parts of the population that best describe the core customer base of the company.

III. Supervised Learning Model: In this part we need to build machine learning model using the result of marketing campaign

and use model to predict which individuals are most likely to convert into becoming customers for the company.I will use several machine learning classifiers and choose the best using analysis of learning curve.

IV. Kaggle Competition: The results of this part needs to be submitted for the Kaggle competition

Datasets and Inputs

There are four datasets, all of which have identical demographics features

- ***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).

In addition to the above data, there are two additional meta-data:

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

Solution Statement

By analysing the attributes of the existing clients and then matching it against a bigger dataset that includes attributes for people in Germany and then essentially figure out which people in Germany are most likely new customers for the client.

i. use unsupervised learning techniques to perform customer segmentation, identifying the parts of the population that best describe the core customer base of the company.

ii. apply what was learned on a third dataset with demographics information for targets of a marketing campaign for the company

iii. Apply supervised learning and use a model to predict which individuals are most likely to convert into becoming customers for the company.

Benchmark Model

Because this is a classification problem, the benchmark model will be the random forest model. According to this article https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2264-5 , the random forest model performs better for large datasets

Evaluation Metrics

According to kaggle, the evaluation metric for this competition is AUC for the ROC curve, relative to the detection of customers from the mail campaign. A ROC, or receiver operating characteristic, is a graphic used to plot the true positive rate (TPR, proportion of actual customers that are labeled as so) against the false positive rate (FPR, proportion of non-customers labeled as customers).

Project Design

Steps:

I. Look at the data to see the shape and how it is are formatted.

II. Data Engineering and Data Cleaning

III. Perform dimensionality reduction and customer segmentation using unsupervised learning techniques.

IV. Make inference and apply what was learned on the third dataset with demographics information for targets of the marketing campaign for the company

V. Apply supervised learning and use a model to predict which individuals are most likely to convert into becoming customers for the company. I plan to compare different models. Because this is a classification problem, a few approaches would be regression, decision trees, SVM, and random forest.

VI. Submit predictions to Kaggle