Machine Learning Engineer Nanodegree

Capstone Proposal

Yau Chi Pui
5th February, 2020

Proposal

Domain Background

Nowadays, almost every company wants to find out their potential customers in order to convert them to the company's customers efficiently. Especially for some mail-order sales companies. The reason is that most people do not respond to the mail, identifying which groups of population have the greatest potential to become company's customers is essential. It can help those companies to allocate resources and boost the sales significantly. This can be done by establishing a customer segmentation model to find the characteristics of the existing customers and general population. Then, we can use supervised learning model to predict the relative probability of an individual to be converted into becoming customers for the companies.

As there are some data provided by Bertelsmann Arvato Analytics, about the demographics data for customers of a mail-order sales company in Germany, demographics information for the general population, and demographics data for individuals who were targets of a marketing campaign. We can use these data to build the models and make predictions which are also applicable to other companies.

Problem Statement

As the general population is tremendous, and most of the individuals do not respond to the mail-sales, it is a challenge for those mail-order sales companies to target potential customers effectively. For instance, only 532 individuals responded to the mail over 42,962 individuals according to the data set provided, which is about only 1.24%. Almost all mail-order sales companies are facing the same problems.

Datasets and Inputs

The datasets used in this project are all provided by Bertelsmann Arvato Analytics, including the demographics data for customers of a mail-order sales company in Germany, demographics information for the general population, and demographics data for individuals who were targets of a marketing campaign.

- azdias.csv: Demographics data for the general population of Germany; 891 211 persons (rows) x 366 features (columns).
- customers.csv: Demographics data for customers of a mail-order company; 191 652 persons (rows) x 369 features (columns).
- mailout_train.csv: Demographics data for individuals who were targets of a marketing campaign; 42 982 persons (rows) x 367 (columns).
- mailout_test.csv: Demographics data for individuals who were targets of a marketing campaign; 42 833 persons (rows) x 366 (columns).

The first two datasets will be used in customer segmentation. Each row of the demographics files represents a single person, but also includes information outside of individuals, including information about their household, building, and neighborhood. The second dataset contains three extra columns ('CUSTOMER_GROUP', 'ONLINE_PURCHASE', and 'PRODUCT_GROUP'), which provide broad information about the customers depicted in the file. We will first extract useful features and then use unsupervised learning techniques to describe the relationship between the demographics of the company's existing customers and the general population of Germany. We can then 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 last dataset will be used in building a supervised prediction model. The dataset will be splitted into two subsets. One will be turned into ‘TRAIN’ subset including one additional column, ‘RESPONSE’, which indicated whether or not each recipient became a customer of the company. The other one, ‘TEST’ subset, the ‘RESPONSE’ column has been removed, will be used to make final predictions that will be assessed in the Kaggle competition.

Solution Statement

We will first extract useful features from the first two datasets by using Principal Component Analysis (PCA), which can retain the ‘principal components’ of the features. Next, we will use these features to train a k-means model that segment the general population to different groups and find out which parts of the population are more likely to be part of the mail-order company’s main customer base. Last, we will use the ‘TRAIN’ subset to train a supervised model, a neural network, to predict the ‘TEST’ subset. This solution is applicable to other companies in this domain if they have enough data.

Benchmark Model

There is a Kaggle competition that is exactly doing the same thing. So, the domain, problem statement, and intended solution are all related to our project. We will use the Kaggle competition to test our model first. Then, we will use the scoreboard of the Kaggle top performers as a benchmark. The top Kaggle performers have approximately 0.8 scores. Therefore, our model should at least have similar score.

Evaluation Metrics

We will first use F1 score to evaluate the solution model as there is a large output class imbalance. We cannot simply use accuracy to evaluate the model. In order to maxmize recall, precision and accuracy score at the same time, F1 score is one of the most appropriate score to evaluate the model. Then, we will upload the result to the Kaggle competition. It will use accuracy to evaluate the model as improving accuracy is our final goal.

Project Design

To sum up, first we will clean the first two datasets so that we can use them in Principal Component Analysis (PCA). We will use PCA to reduce the feature dimensional and extract most useful features. K-mean clustering will be applied later by using the above extracted features so that we can group the general population into different segments. Then, we will find out which groups are more likely to be part of the mail-order company’s main customer base and extract those groups’ features. Those extracted features and the third dataset will be used to train a Neural Network model to predict the ‘TEST’ subset. We will use F1 score to evaluate the model first, then we will upload the result to the Kaggle competition to check the accuracy and use the scoreboard as a benchmark of our model.

Data repository: https://github.com/Vendetta37/Capstone-Project/tree/master/source