Machine Learning Engineer Nanodegree

Capstone Proposal

Vagner A Silva

August 21st, 2020

Proposal

Domain Background

The dataset is a mail-order sales company in Germany (Arvato Financial Solutions, a Bertelsmann subsidiary) and is interested in identifying segments of the general population to target with their marketing in order to grow. Demographics information has been provided for both the general population at large as well as for prior customers of the mail-order company in order to build a model of the customer base of the company. The target dataset contains demographics information for targets of a mailout marketing campaign.

The objective is to identify which individuals are most likely to respond to the campaign and become customers of the mail-order company.

Problem Statement

To resolve this problem is necessary for the first to see the customer's dataset using unsupervised learning algorithms.

Whit this they propose is to identify segments in general in the existing customers

The next step is to apply one supervised learning algorithm will be used to make predictions on whether a person is a probable customer or not, based on the demographic data.

Datasets and Inputs

All the datasets were provided by Arvato in the context of the Udacity Machine Learning Engineer Nanodegree.

There are 4 datasets to be explored in 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).

And 2 metadata files associated with these datasets:

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

Solution Statement

The first step to do is, the task is to identify any customer segments present in the provided dataset and Match these segments with the segments of the population present in the dataset. Including explore to examine if there are any missing values or misrecorded values in the data and fix them and verify and made adjustment features to no have higher weights.

The next step is to identify the minimum number of features, in a dataset like these many features are not representative. To do this we apply PCA (Principal Component Analysis).

And the final step is to segment the customers into different segments based on the selected features, to do this we will apply an unsupervised learning algorithm K-means clustering is a good choice for this.

Following the project, entering into the task to predict can acquire a new customer

In the second part of the project, the task is to predict whether the direct mail company can acquire a customer.

In the first step, the data is pre-processed (the first two steps of the first part will be performed again on the train and test data) and the next supervised learning algorithm will be trained and evaluated on the pre-processed training data.

To finally step the trained model will be used to make predictions about the test data provided.

Benchmark Model

For this dataset and this problem, one benchmark model, would be a Logistic Regression model since it is easy to train and test within less amount of time.

The performance of this model will be considered as a baseline for further, where different algorithms can be used to compare the performance with this benchmark to decide whether to proceed with a good algorithm.

Evaluation Metrics

In the Supervised Learning section, we had to pick a metric in order to evaluate different models. Since the data is imbalanced, with ~1% of users responding positively to the campaign, AUC/ROC is used as the success metric for this part, rather than accuracy.

The evaluation metrics most common for classification:

- Accuracy
- Confusion Matrix – F1 score, Recall, Precision
- Area Under the Receiver Operating Curve (AUROC)

One big important target is the imbalance to the dataset if happened the confusion matrix is a good idea to evaluate one good classifier.

Project Design

1. Data Cleaning and Visualisation to find data with problems and fixing, analysis yo understand to patterns in the data.

2. Reduction to dimensionality with PCA to find correlations between features the dataset and identify important features with this important analysis using K-means to clustering and find clusters

3. Using different supervised algorithms will be trained and evaluated in the context of predicting whether a person will be our next customer or not. Algorithms like Logistic Regression, Decision Tree, Random Forests and Gradient Boosted Trees will be may be used to make predictions and will be evaluated. Obviously, the better model will is tuning and used.

4. Finally, the best model will be used to make predictions on the test data and the predictions will be submitted to the report this project.

Links

[1] https://www.kaggle.com/c/udacity-arvato-identify-customers/overview/description