Machine Learning Engineer Nanodegree Capstone Proposal

Jing Han Ng
December 24th, 2020

Domain Background

Bertelsmann Arvato analyses attributes of customers to predict individuals which are most likely to be a potential customer for a company. The project will aim to do the same. By predicting the attributes of individuals most likely to be a future customer for a company, based on existing data and attributes of existing customers, the company can reduce marketing expenditure and increase sales. The problem of predicting a likely future customer is a solvable one as we have existing demographics data of current customers to build a prediction model on. Anecdotal evidence that this is possible is offered by Bertelsmann Arvato itself which have built a business around such predictive analytics. Academic research evidence by the University of Western Ontario has shown that such problems can and have been solved many years ago[1]. The result of the research proved that predicting customer response via data mining improves profits in comparison to mass marketing.

Problem Statement

The problem to be solved involves predicting who will respond to a mail-out campaign, based on the analysis of demographics data available. The analysis will be performed using unsupervised and supervised machine learning algorithm. This problem involves an imbalanced data set with very low occurrences of true positives.

Hence the results of the model will be measured based on Area Under the Curve(AUC) of Receiver Operating Characteristics(ROC).

Datasets and Inputs

The dataset provided by Arvato Financial Solutions in this problem are as follows:

1.) Udacity_AZDIAS_052018.csv - Demographics data for the general population of Germany. The dataset comprises 891211 persons (rows) x 366 features (columns).

2.) Udacity_CUSTOMERS_052018.csv - Demographics data for customers of a mail-order company. The dataset comprises 191652 persons (rows) x 369 features (columns).

3.) Udacity_MAILOUT_052018_TRAIN.csv - Demographics data for individuals who were targets of a marketing campaign. The dataset comprises 42982 persons (rows) x 367 (columns)

4.) Udacity_MAILOUT_052018_TEST.csv - Demographics data for individuals who were targets of a marketing campaign. The dataset comprises 42833 persons (rows) x 366 (columns). This data set has the RESPONSE column removed to allow for our model prediction to be assessed against.

There are an additional 2 files providing description of the features (columns) of the data sets:

1.) DIAS Attributes - Values 2017.xlsx

2.) DIAS Information Levels - Attributes 2017.xlsx

Solution Statement

The solution to the problem can be quantified based on area under curve of the receiver operating characters, AUC of ROC. We shall aim to optimise the value of AUC of ROC. The AUC of ROC signifies the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. This metric is applicable as our data of true responses is very low. Because of this, we cannot evaluate our model based on other metrics such as accuracy as the model will bias towards predicting a non-response from customers; this will produce a “high accuracy” model which will be erroneous - as the model is only accurate due to the the low occurrence of true positive responses. A solution based on the result of AUC of ROC is hence applicable to this project, appropriate for the dataset, quantifiable and replicable.

Benchmark Model

The classifier model used are Gradient Boosting Classifier, Adaboost Classifier, and Random Forest Classifier. The model will be selected based on the AUC of ROC and evaluated by using the learning curve method to determine if additional input data might improve result.

Evaluation Metrics

The evaluation metric that is proposed is AUC of the ROC. A ROC is a graphical plot obtained by plotting the true positive rate (TPR,) vs the false positive rate (FPR) across all possible classification threshold. Since this problem involves a dataset which is imbalanced, AUC of the ROC is the appropriate solution.

Project Design

The workflow for completing the project is as below:

<u>Preprocessing</u>

- Cleaning of the dataset by checking that data with null and missing values are removed.
- Cleaning of dataset by dropping columns and rows with many missing values.
- Cleaning of the dataset by feature classification and engineering of dataset.

- Cleaning of the dataset by checking for outliers and dealing with them.

Part 1: Customer Segmentation Report (Unsupervised Learning)

- Reducing number of columns/features with Principal Component Analysis (PCA).

- Application of K-Means clustering algorithm to obtain the attributes of current customers and their difference to the general population.

Part 2: Supervised Learning Model

- Build a prediction model with the ensemble learning method such as Random Forest Classifier, Adaboost Classifier and Gradient Boosting Classifier.

- Perform model tuning using Grid Search

Part 3: Kaggle Competition

- Evaluate and test the best model after parameter tuning by submitting the result to Kaggle

References

1.) C.X. Ling, C. Li, “Data Mining for Direct Marketing-Specific Problems and Solutions”, Proceedings of the Fourth International Conference on Knowledge Discover and Data Mining (KDD ’98), pg. 73-79, 1999