Capstone Project Proposal

Machine Learning Engineer Nanodegree

Vinicius Gomes Pereira

29/02/2020

The Market Segmentation Problem

Market segmentation consists of identifying a group of individuals who have characteristics in common, in a heterogeneous market. This should be seen as a powerful tool for auxiliary departments of marketing, design and strategic areas.

Several characteristics of consumers and potential consumers can be taken into account for this segmentation task, for example:

1. Demographic, geographic, social and economic criteria.
2. Personality and lifestyle criteria.
3. Product behavior criteria

The question is what is the advantage of segmentation? The market is heterogeneous and therefore it is advisable to segment it. Thus, attitudes towards these segments can be different. Segmentation allows to rationalize the means to reach a given product segment, adjusting it to the prices and costs of distribution and communication, with a view to achieving balance. It also allows a specialization of the company playing with the strategic variables - price, product, distribution and communication - avoiding waste.

Problem Statement

Consider a company's marketing campaign (Arvato Financial Services), in which we need to select those individuals who can become the company's future customers. For this task, we have the following databases: demographic information from Germany (country where the company is located) and information from individuals who are already customers of this company.

We used unsupervised machine learning algorithms to analyze the market and segment it, selecting the main characteristics that can best describe the company's consumer.

Then, we used this information to create a predictive model that can determine with reasonable precision whether an individual can be a likely consumer or not, when subjected to a particular marketing campaign.

In addition, we developed a web service so that, when introducing characteristics of a certain person, the service responds if the individual will become a potential consumer.

Datasets and Inputs

There are four data files associated with this project, that were provided by Udacity :

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).

The first and the second file will be used in the customer segmentation step. The third one will be used to train the predictive model that will be used to identify possible consumers. The forth file will be used to measure how well our model performs in the classification task.

Solution Statement

For Customer Segmentation Report, we are going to do the exploratory data analysis on both datasets, apply unsupervised machine learning techniques like K-means with some dimensionality reduction method (PCA, for example), extracting important features from the data.

After that for the Classification Task, we are going to use the extracted features from the prior analysis and apply some classification algorithms (like Random Forest, SVM, Logistic Regression, Decision Trees and XgBoost).

Finally, the developed algorithm will be applied to the test dataset and submitted on Kaggle's platform. Then, we are going to deploy on AWS and make the model available for using through a web service.

Benchmark Model

First we are going to compare, in the classification task, our developed model, using the features that we extracted, to Naive Bayes Model using the original features of the dataset. Then, we are going to compare our model to the Kaggle ranking of this project [1].

Evaluation Metrics

Accuracy, recall and precision are common metrics to use for classification task. These three evaluation metrics are defined below, where TP, TN, FP, FN mean true positives, true negatives, false positive and false negatives in the classification process:

accuracy = (TP+ TN)/(TP + FP + TN+FN)
recall = (TP)/(TP+FN)
precision = (TP)/(TP+FP)

Besides, we are going to also evaluate the AUC for the ROC curve defined in [2]. The ROC curve is a graphic used to plot the true positive rate TPR (true positive rate) against FPR (false positive rate).

Project Design

We are going to follow the next steps:

1) Exploratory Data Analysis of demographics data for the general population of Germany;
2) Exploratory Data Analysis for the consumers dataset;
3) Feature Extraction, using Principal Component Analysis (PCA);
4) K-Means for customer segmentation, using the features extracted;
5) Develop classification models on training dataset, using hyperparameter tuning and validation techniques,

6) Choose the best model and submit on Kaggle platform;
7) Deploy the chosen model on AWS SageMaker;
8) Enable an endpoint for the the Web Service, using (API Gateway on AWS and Lambda Functions);

References

[1] Kaggle.com. (2020). *Udacity+Arvato: Identify Customer Segments* | *Kaggle*. [online] Available at: https://www.kaggle.com/c/udacity-arvato-identify-customers/leaderboard [Accessed 1 Mar. 2020].

[2] En.wikipedia.org. (2020). *Receiver operating characteristic*. [online] Available at: https://en.wikipedia.org/wiki/Receiver_operating_characteristic#Area_under_the_curve
[Accessed 1 Mar. 2020].