Machine Learning Engineer Nanodegree

Udacity

Capstone Project Proposal:

Customer Segmentation (Analysis)

Miguel Diaz

February 21st, 2020

Table of contents

I. Domain Background

Bertelsmann was found as a publishing house in 1835 by Carl Bertelsmann (Stephan Grimm, 2012), in Gutersloh, North Rhine-Westphalia, Germany. Since then, the company has evolved and developed different services to become a private multinational conglomerate.

Arvato is one of the eight divisions, that its services include customer support, information technology, logistics, and finance. The financial services include a wide range of solutions that include payment processing, factoring, and debt collection services. Besides, the company's credit rating and risk management activities, as well as insurance services, are offered in this field (Andreas Toller, 2016).

We are proposed to find the best solution to acquire new client base with the datasets provided and Machine Learning techniques. It is important to plot the data make some modifications, and model it in order to segment interesting customers and resolve this task.

II. Problem Statement

The problem statement is no other than finding the best way the company can acquires new client base in an efficient way but is important also to have a solid business background to save time and sources.

III. Datasets and inputs

The data is contained in 4 datasets and 2 metadata files:

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 the 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).
DIAS Information Levels – Attributes 2017.xlx: Top-level list of attributes and descriptions, organized by informational category.

DIAS Attributes – Values 2017.xlsx: Detailed mapping data values for each feature in alphabetical order.

IV. Solution Statement

The problem stated before proposes to apply Machine Learning techniques to predict good recipients by analyzing data of specific populations and their relationships between them.

We need to use unsupervised and supervised machine learning techniques, but before we apply them, we have to analyze the best possible algorithms for the specific purposes of this problem.

For the first task we need to clean and apply techniques for unsupervised modeling, for example: encoding data, feature scaling and handling missing data in order to not affect the results. We will use PCA for dimensionality reduction and identify variables (principal components) whose behavior as a function of simulation parameters can expose the presence of phase transition; and a K-Means algorithm as a way of splitting up the cluster and as part of the prediction process. Besides being one of the most popular among clustering algorithms, it has a high dependency on the initial conditions so the implementation could change according to the size of the dataset. We could implement recursive and parallel approximation to this algorithm in order to scale well on both the number of instances and dimensionality of the problem.

When we get the cluster segments, we will apply the supervised learning models to test them and analyze their final results to compare them and choose the best approach. We will use:

a. Logistic Regression
b. XGBoostClassifier (DecisionTree Regressor)

As Machine Learning engineers, we sometimes need to adjust our approaches to get valuable insights and the best possible results, so we need to be flexible and try different experiments.

V. Benchmark Model

According to the Machine Learning literacy, Logistic Classification algorithm is easier to implement and make no assumptions about distributions of classes in feature space, another characteristic that determines this algorithm is that the procedure for determining (training) the optimal weights/coefficients and the way that the score is interpreted. Here the score can directly be converted to a probability value, indicating the probability of observation i choosing outcome k given the measured characteristics of the observation.

The XGBoost Classifier algorithm has been consistently placing among the top contenders in Kaggle competitions. It implements the stochastic gradient boosting and offer a wide range of hyperparameters, and it is focused only on decision trees as base classifiers, a variation of the loss function is used to control the complexity of the trees.

VI. Evaluation Metrics

What we want to solve is the problem of customer segmentation, by applying machine learning techniques on demographics and customer data.

At first place, we implement a K-Means, because we need to reduce high-dimensional data set into fewer dimensions while retaining significant data, allowing to untangle data into independent components. So, it would be possible to identify the segments.

For the second step, in the supervised learning modeling, it’s important to use precision, recall, accuracy, F1, AUC

as main metrics to evaluate the model performance and customer classifications.

The outputs of this analysis are part of a marketing strategy to identify potential clients.

VII. Project Design

a. Data loading

Load files with the specific requirements.

b. Data Exploration (Visualization)

Visualize data to understand the patterns and distributions.

c. Data pre-processing (Cleanup)

Fix issues like missing values, outliers, feature types, etc.

d. Data Transform (Feature engineering)

Prepare data to meet the problem statements like encoding data, scaling data, or creating new features.

e. Model Selection

Experiment with different algorithms to find the best for our problem.

f. Model Tuning

Adjust the model parameters for increased performance without overfitting or underfitting.

g. Data Splitting (Train, validation, test)

The largest dataset is for training the algorithm, the validation set, to evaluate the model and the test set, to measure the final results of the process.

VIII. References

[1] R. M. Woloshyn, “Learning phase transitions: comparing PCA and SV M”, ArXiv, 2019.

[2] M. Capo, A. Perez, and J. A. Lozano, “An efficient K-means clustering algorithms for massive data”, ArXiv, vol. 14, no. 8, 2015.

[3] C.P. Ezenkwu, S. Ozuomba, C. Kalu, “Application of K-means algorithm for efficient customer segmentation: A strategy for targeted customer services”, IJARAI, vol. 4, No. 10, 2015.

[4] P.C. Chaitra, K. Saravana, “A review of multi-class classification algorithms”, IJPAM, vol. 118, No. 14, 2018

[5] C. Bentejac, A. Csorgo, G. Martinez-Munoz, “A comparative analysis of XGBoost”, ArXiv, 2019.