Create a Customer Segmentation Report for Arvato Financial Services

Pedro Szloma Herr Zaterka

Abstract—Customer behavior and marketing campaign efficiency are a common topic in business, and a prosperous field for machine learning practitioners. As an example, the problem tackled is both a customer segmentation and marketing response problem. This problem is presented in two steps: A clustering problem (Unsupervised Learning) and a classification problem (Supervised Learning). This capstone proposal presents the problem, data to be used and approach to the problem. Techniques intented to be used vary from standard logistic regression models to complex neural networks.

Index Terms—Machine Learning, Customer Segmentation, Marketing, Udacity, Arvato, Kaggle.

1 Introduction

The proposal presented here is one of the 3 suggested by the Machine Learning Engineer Nanodegree, the goal of it is to analyse demographics data for customers of a mail-order sales company form Germany, to perform customer segmentation using unsupervised learning to identify what customer profile in the general population would be more likely to become new customers.

After that, using a mail-out dataset, a new model will be created to predict whether or not a customer who receives the mail will convert or not. This task is also part of a Kaggle competition [1] and its results will be submitted to it as well.

2 Domain Background

The project aims to be a consulting project in marketing and sales. The premises are that patterns can be found within the client database that can be extrapolated to the general public, making it able to identify new potential customers, as well as using this pattern to predict how likely someone is to convert after receiving a marketing campaign.

It is a very broad and generic problem in terms of marketing campaigns and its a common ground in many business. Both using clustering methods to understand the customer profile as well as knowing the likelihood of your own client database to have a positive response to a marketing campaign.

The reasoning behind choosing this problem is that its a very generic and common problem in data science, as I have some experience in the area but have yet to build a portfolio example, this will fit nicely as well as encouraging me to enter in the Kaggle environment, which is also a personal goal I have.

3 Problem Statement

The problem consists of 3 parts:

- Perform a unsupervised learning task to identify patterns in customer data and compare it to Germany's population and create a customer segmentation report
- Perform a supervised learning task to predict likelihood of people to respond positively yo a campaign
- Use the second model in a Kaggle competition

4 Datasets and Inputs

There are 3 datasets in this project:

1) Clustering part:
   a) Demographics data for the general population of Germany
   b) Demographics data for customers of a mail-order company
2) Classification part:
   a) Demographics data for individuals who were targets of a marketing campaign

As well as two aditional files ”DIAS Atributes” and ”DIAS Information Level” that provide aditional information regarding the features in the datasets. All data can be found in the Udacity Workspace.

5 Solution Statement

For the unsupervised learning task, the initial approach will be using a technique like Principal Component Analysis (PCA) to reduce dimensionality and apply a clustering method (like K-Means Clustering), then analyse the main components for the clustering and what features have a strong influence in them.

For the supervised learning, firstly simpler models like regression and random forest will be tested alongside XGBoost [2] and LightGBM [3]. After that, feature extraction, pre-processing methods and neural networks will be continuously implemented to try achieve higher scores, in conjunction with hyperparameter optimization methods such as the Tree-structure Parzen Estimator Approach, to achieve the best hyperparameters possible for the final model.

6 Benchmark Model

As there is the public leaderboard of the Kaggle competition, the benchmark will be the top 10 AUC ROC scores achieved by the other competitors (0.808 at the time of this writting).

7 Evaluation Metrics

The evaluation metric for the supervised model will AUC for the ROC curve, as stated in the competition. Unsupervised metric is more subjective, and it will depend on the exploratory analysis to be performed.

8 Project Design

The first step is to perform an exploratory data analysis, to see relations between features, missing values, possible outliers and noise in the data. This will also allow for the possible removal of features of access the necessity of encoding any categorical columns.

This exploratory data analysis will also raise some hypothesis about the data, especially when comparing statistic properties of some features from the customer database to the general german public.

After the early pre-processing and data cleaning done on the datasets for the demographics dataset for customers of the mail-order company (dataset for the unsupervised task), a PCA will be performed to reduce dimensionality, the number of components will be tried until a fitting number is achieved, and then the clustering method will be applied, first trying K-Means Clustering.

When the clustering model is done and patterns are found, the criteria will be applied to the German population database to find what portion of the database would be potential clients.

With the clustering done, the supervised learning task begins. As it happened in the previous task, an exploratory data analysis will be carried on in the marketing campaign dataset, to find out possible patterns, outliers and correlation between features. If needed, dimesionality reduction will be applied, firstly with PCA but a Convolutional Neural Network can also be used to find patterns in the data and provide an intermediate output to feed other models.

For the machine learning model, it should start with simpler models, such as logistic regression and random forest classifier (the later will also be used for feature selection), than going to boosting algorithms such as XGBoost and LightGBM. Neural Networks shall also be tested, with architectures based on a mixture of convolutional with fully connected layers, as well as the TabNet [4] architecture.

The best models found will have their hyperparameters tuned with the Optuna [5] package (which from previous experiences works a little better than HyperOpt and Scikit Optimize) and their perfomance will be measure in the Kaggle competition.

References

[1] "Kaggle competition: Udacity+arvato: Identify customer segments." https://www.kaggle.com/c/udacity-arvato-identify-customers. Accessed: 2021-01-25.

[2] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), pp. 785–794, ACM, 2016.

[3] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, pp. 3146–3154, Curran Associates, Inc., 2017.

[4] S. O. Arik and T. Pfister, "Tabnet: Attentive interpretable tabular learning," *arXiv preprint arXiv:1908.07442*, 2019.

[5] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.