Capstone project proposal

Customer Segmentation – Arvato Financial Solutions

Job Isaias Quiroz Mercado

February 2021

Background

This project was proposed by Arvato Financial services. Arvato Bertelsmann is a company dedicated to help its customers with several financial aspects of organizations. Such aspects include invoicing & accounting, debt collection, credit risk & fraud management, among others. Its solutions have been applied in industries ranging from e-commerce to telecommunications and banking.

Marketing is essential for any sales-related business; however, marketing campaigns are expensive and sometimes even risky. In this project we contribute to create an effective marketing campaign that uses machine learning techniques to estimate which individuals can be potential clients (according to some data). The final solution can help to reduce the cost of marketing while increasing the chance of having new customers.

Problem statement

From a business perspective, the problem statement of this project is to answer the following question: "How can our client (a mail-order company of organic products) acquire new clients more efficiently?"

Datasets and inputs

Arvato has provided four datasets with attributes and demographic information from existing clients, and from a sample of the general population in Germany.

- Udacity_AZDIAS_052018.cscv: Demographics data for the general population of Germany; 891,211 subjects with 366 features each.
- Udacity_CUSTOMERS_052018.csv: Demographics data from past customers of the mail-order company; 191,654 subjects with 369 features each.
- Udacity_MAILOUT_052018_TRAIN.csv: Train subset of targets individuals of a marketing campaign; 42,982 subjects with 367 features each.
- Udacity_MAILOUT_052018_TEST.csv: Test subset of targets individuals of a marketing campaign; 42,982 subjects with 367 features each.

Arvato has also provided two additional files describing the meaning of the features in the above datasets.

- DIAS Information levels

- DIAS Attributes

Solution statement

The solution is straightforward: based on the features of a person's features, we have to develop a model to predict whether or not such person has a higher chance to become a customer of the mail-order company. However, since the number of subjects and features is high and, since we have additional useful information (from the existing clients), we can first investigate which features are more useful to predict our goal.

The first part of the solution is to use an unsupervised learning algorithm to match the features of the existing clients with those from the general population in Germany, the goal of this is to characterize the profile of potential customers.

As a second step, we can proceed to train a classification model to predict whether or not a person has a high possibility of becoming a client.

In the first case, a convenient algorithm to try is KMeans, because it is an efficient model and performs well with medium size datasets. If necessary, we can use a dimensionality reduction technique (such as PCA) to improve the performance of the model. In the second case, it can be useful to try with a simple linear model (such as logistic regression) and also try with a more complex algorithm such as a RandomForest classifier.

Benchmark models

There is not an specific benchmark model for this project. However, the associated Kaggle competition indicates that the Top 80 participants reach a AUROCC score above 0.8.

Evaluation Metrics

The confusion matrix is always worth computing, but the most relevant metrics will be Area Under the ROC curve (AUROCC). This metric can be interpreted as "the probability that the model ranks a random positive example more highly than a random negative example" [2].

Project design

1) Data cleaning. The first step of any data-related project is to start knowing the datasets while getting rid of unnecessary information.
2) Feature engineering and population segmentation. Since the number of features is reasonably high, it might be useful to apply PCA to reduce the dimensionality of the examples. In this step, we want to identify groups of subjects that are potential customers.

3) Model selection and training. As stated above we can start by training a linear model and see how it performs on the training set. Depending on the results we can consider more complex non-linear algorithms such as the RandomForest classifier.
4) Model tunning. Once we have selected one or two models we can proceed to fine tune its hyperparameters. Depending on the number of hyperparameters we can opt to use GridSearch or a non-exhaustive way to tune the hyperparameters.
5) Testing. We use the test set to evaluate and improve our model.

References

[1] Arvato Bertelsmann, “About us”, https://finance.arvato.com/en/about-us/

[2] Google, “Classification: ROC Curve and AUC”. Available on: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc