Project proposal: Customer Segmentation Report for Arvato Financial Services

Robert Kwapich, June 2020

Domain Background

Project focuses on the customer segmentation, an analysis of demographic data of customers of a Arvato company.

Arvato is a mail order sales company in Germany (also in other countries) that wants to leverage information about the general population in Germany to identify parts of this population that are most likely to use or be persuaded to use their services.

Moreover, the company has gathered historical data from the past marketing campaigns that recorded demographic info, as well as information about success or failure in convincing a person to become Arvato's customer.

Arvato kindly provided anonymized data on their customers, the results of the marketing campaign and the general population demographics of Germany (see **Datasets and Inputs** below for details)

Problem Statement

The main problem of the project is to provide a solution for Arvato problem of acquiring new clients in a more efficient way. The problem lies in utilizing their existing data: client information as well as past marketing campaigns, along with dataset about general population. Arvato would like to understand their clients more, with regard to general population and their likelihood of potential use of service.

In particular, the project consists of serval problems to solve:

1. **Exploratory Data Analysis (EDA) and data preparation**

Provided data should be explored, standardized, cleaned and pre-processed in a way that would enable subsequent unsupervised and supervised analysis possible: providing information about possible models.

The extent to which missing data are present should be analyzed, and appropriate steps should be taken: removal or imputation.

2. **Customer data segmentation**

Quantify and describe the relationship between the demographics of Arvato's customers and general population of Germany.

The goal is to extract valuable information about general population of Germany, and identify segments that are more likely to become part of Arvato client base.

This data could be then used for targeted marketing campaigns that could present Arvato offer to the people that are most likely to use it.

3. **Customer aquisition prediction**

Utilizing previously explored customer demographic data, with Arvato's marketing campaign results, it is possible to create and deploy a supervised model for classification. This model will provide a classification of whether a customer will likely or won't become a Arvato's client in the future.

Serveral different model architectures will be used, each of which will be tuned with respect to the optimal hyperparameters.

4. **Kaggle competition / Model Validation**

Provided **test** dataset purposefully doesn't contain the prediction labels, so as to objectively validate the final model performance for all participants in the **Kaggle** data science competition: link here.

The objective here is to submit the prediction results for the best obtained model, and obtain the final AUC score.

Datasets and Inputs

The datasets were provided by Arvato company in a non-open disclosure. That is, the dataset is not open for the public, and is made available for this analysis only. Full terms and contions are described in an appropriate file `terms.pdf` in this this directory.

The main dataset comprises of four files:

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).

In addition two metadata files with descriptions are provided:

1. `DIAS Information Levels - Attributes 2017.xlsx` : a top-level list of attributes and descriptions, organized by informational category.
2. `DIAS Attributes - Values 2017.xlsx` : a detailed mapping of data values for each feature in alphabetical order.

Solution Statement

As described in **Problem statement** section, the problem at hand is divided into serval steps. In this section a potential solution is discussed.

1. Exploratory Data Analysis.

Provided features should be encoded, normalized and standardized (i.e. scaled). Missing values should be tackled in one of potential ways: removal or imputation. Finally, feature selection will be performed for subsequent analysis.

2. Unsupervised analysis.

Using Principal Component Analysis (PCA) for dimensionality reduction, with subsequent `K-means` clustering, with the "elbow-method" to identify optimal number of clusters.

3. Supervised analysis.

Supervised portion of the analysis needs to learn a function mapping customer features into a binary (or probabilistic) measure for decision indicating whether a particular client is likely to respond positively to a targeted marketing campaign, i.e. become a client in the future.

For this purpose several models could be tested:

```
- **linear regressor**: a baseline, and most simplistic model, used for comparisons with more complex ones,
- **tree boosting**: (`XGboost` alrogithm): one of the most-widely used method for regression/classification in machine-learning, re
- **Neural network**: a Pytorch/Tensorflow deep neural network. The aim is to utilize non-linear classification capabilities of thes
```

In addition, a **hyperparameter tuning** will be used on a portion of test data, to find optimal values for various (often arbitrary) hyperparameters used in selected models.

The model that achieved the highest value of `AUC` (Area Under Curve) will be considered the most optimal for the task at hand, and choses as the final solution.

Benchmark Model

As discussed above, the benchmark model is a basic linear regressor, against which two models discussed above will be tested. The goal is to establish a naive baseline with a benchmark model, and improve upon that solution taking into consideration the evaluation metrics (discussed below).

Evaluation Metrics

Since, the project contains several steps, there is a separate metric for each:

1. Unsupervised Learning - PCA.

For PCA analysis used for dimensionality reduction, the amount of **variance explained** is used as a metric for the analysis of features ("feature selection").

2. Supervised Learning

For the supervised learning models metrics such as: accuracy (if the classes are well-balanced), precision/recall, or binary cross-entropy are good first candidates (for binary, probabilistic decisions).

Project Design

Summarizing theoretical workflow:

1. EDA part 1: **Data cleanup**:

Identifying missing or wrong values (i.e. a string in numerical field). Deciding whether to drop a feature in case it has high percentage of missing values, remove certain entries, or attempt to impute missing values.

2. EDA part 2: ** Visualization ** Part of intuition building, getting to know the overall distribution of values, identifying potential problems with dataset, feature correlations, ranges and scales on which they occurr.

3. EDA part 3:**Data standarization and processing**

Identifying which features need to be standardized: z-scored, log transformed, encoded as numbers, etc., and preparing a standardized input dataset used later in the analysis.

Splitting the initial training dataset into: **train** and **validation datasets** for which we'd have true labels. The provided **test** dataset is the final dataset for obtaining final score, but the true labels are not available directly. Jobs such as hyperparameter tuning, and model selection should be evaluated on a separate dataset, which is independent from the training set.

4. Unsupervised analysis:

   a. Feature selection Apply PCA on cleaned-up and preprocessed dataset. Find most relevant features with feature selection: using the amount of variance explained as a criterion, as well as feature correlation matrix ("confusion matrix") to identify redundant features.

   b. K-means clustering with elbow method Find optimal number of cluster though a simple heuristic using "elbow method". In more advanced scenario

Hierarchical Dirichlet Process clustering could be used.

5. Supervised analysis:

Model selection: train previously-mentioned selection of models: linear regressor, tree-boosting (XGBoost) and Neural Network. Perform hyperparameter tuning and model comparison for a selected metric. Choose the best performing model evaluated on a test set.

6. Submit solution to Kaggle competition.

Arvato company created a Kaggle competition into which predicted labels for a **test** dataset need to be submitted. The result is a final score is based on AUC, and more details are described on Kaggle competition website for Arvato project.

7. (Optional) Re-submit solution to Kaggle

If the final result is not satisfactory for the purpose of the analysis, the proposed solution needs to be re-evaluated: re-consider feature selection, number of population clusters to include, variance explained to re-consider, propose a new model architecture, etc... Essentially re-iterate all steps.

Sources:

Below I include some method reference materials for the project: 1. Scikit learn PCA

2. Scikit learn imputation
3. Scikit learn K-means
4. Sagemaker documentation: a. Linear lerner

   b. XGboost

   c. PCA

   d. K-means

   e. Tensorflow estimator

   f. Hyperparameter tuner

Proposal follows a well-organized structure and would be readily understood by its intended audience. Each section is written in a clear, concise and specific manner. Few grammatical and spelling mistakes are present. All resources used and referenced are properly cited.