Arvato-Proposal

December 8, 2019

1 Machine Learning Engineer Nanodegree

1.1 Capstone Proposal

Sebastian Uribe Ocampo

1.2 Domain Background

Arvato Bertelsman (Financial Solutions Division) is an international financial service provider from Germany, they offer solutions related to ID & Fraud Managemet, Credit Risk Management, Payment & Financing Services and Debt Collection Services. Like every company with a marketing department, it is important to understand who their customers are, what are their characteristics, how is their payment behavior and more, all important information to provide a good service, but most importantly to keep profits up !.

For any company(not only Arvato), designing efficient marketing campaign on how to acquire new customers is not an easy task, it involves a series of multiple studies to understand customers information and design strategic campaigns to maximize customer's acquisition taking into consideration the cost of the campaign. You can't just send ads to any non-client with publicity, too much money spend and with little ROI 1, its more effective to send ads to those who shares similar characteristics to your already clients. The big and important questions is "*what are those characteristics that groups my clients based on their characteristics ?*". Thankfully some Machine learning methods we can answer it.

1.3 Problem Statement

The goal of this project is to perform and analyze a customer segmentation that allows the principal characteristics of the core business customers. This finding's will be later be used as part of information to predict which individuals are most likely to convert into new customers.

From Arvatos perspective, this project might help answering the following questions:

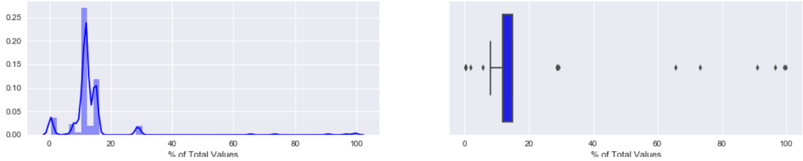
- *How can the mail-order company acquire new clients more efficiently.*
- *Which people are most likely new customers ?*

This kind of problem is both an unsupervised and supervised machine learning problem. For the unsupervised part of the project, we are provided with features of demographic data from population of Germany and customers of the mail-order company to apply any clustering technic looking patterns to segment posible clients. There is not right anwser here, but the answer must be according to the financial needs of Arvato. For the supervised part, it is clearly a classification problem, with a binary target that represents which individuals are most likely to be new customers (Represented as **1**) and those who are not very liekly (Represented as **0**).

1.4 Datasets and Inputs

There are four data files associated with this project:

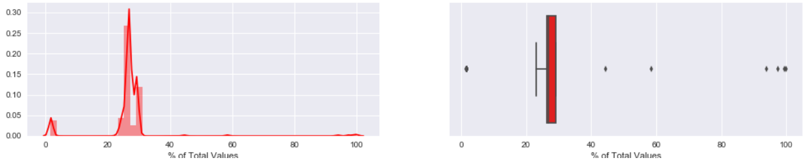
- Udacity_AZDIAS_052018.csv: Demographics data for the general population of Germany; 891211 persons (rows) x 366 features (columns); There are 273 features with at least one missing value.



Missing values azdias dataset

In a quick exploration, it is seen that are a few columns where almost all data point are *null*. Mainly features are missing between 10-15 % of the data.

- Udacity_CUSTOMERS_052018.csv: Demographics data for customers of a mail-order company; 191652 persons (rows) x 369 features (columns).There are 273 features with at least one missing value.



Missing values customer dataset

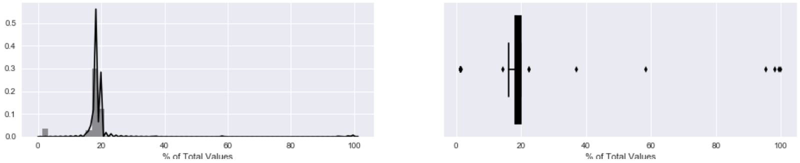
Most feutures miss around 25 % of the data.

- Udacity_MAILOUT_052018_TEST.csv: Demographics data for individuals who were targets of a marketing campaign; 42833 persons (rows) x 366 (columns). There are 273 features with at least one missing value.

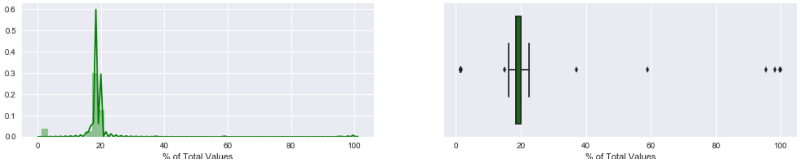
Most feutures miss between 15-20 % of the data.

- Udacity_MAILOUT_052018_TRAIN.csv: Demographics data for individuals who were targets of a marketing campaign; 42982 persons (rows) x 367 (columns). There are 273 features with at least one missing value.

Most feutures miss around 20 % of the data.
Checking just the response feature -> This is a highly imbalanced dataset!.



Missing values test



Missing values train dataset

1.5 Solution Statement

First I will use the DIAS attributes file to get an idea on how to clean the data, there are a lot of missing data, so this part might be quiet long since there are numeric and categorical features. For the unsupervised section of the project I will use PCA in order to reduce the high dimensions and perform a cluster with the principal components founds that at least explain 80 % of the cumulative variance. Then I will use some clustering technique like K-means using the principal components as features to group the clients into clusters and comparing the resulting clusters with the mail-company clients.

For the Supervised section, I will compare some classifiers like Logistic Regression, Decision Tree, SVM or even some ensemble methods using the preprocess features "X input" to predict the customers " Y output" probability of becoming new customers. An AUC above 0.75 will be considered a successful classifier, since its predictive power its considerable above randomness.

1.6 Evaluation Metrics

Customer Segmentation Stage : Commonly, unsupervised models don't have a right answer, this must depend on the type of problem and the business goals, but we can use the Silhouette Score to get an idea of consistency within the clusters, where the intra-cluster distance (a) and the mean nearest-cluster distance (b) is calculated for each sample.

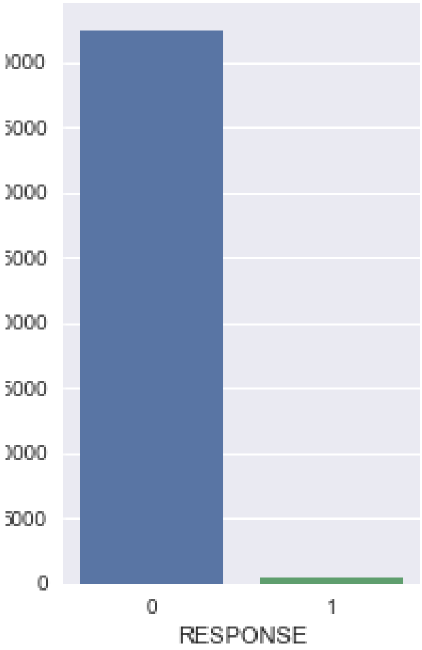
The Silhouette Coefficient for a sample is :

$$SilhouetteCoefficient = \frac{(b-a)}{max(a,b)}.$$

Same again, without knowing much about the business goal we can use the Elbow Method Heuristic to determine the optimal number of clusters by measuring the within cluster sum of errors (*WCSS*). *WCSS* is the summation of the each clusters distance between that specific clusters each points against the cluster centroid.

Supervised Stage :

Area Under the Receiver Operating Characteristic Curve ROC AUC is a common metric used when predicting the probability for binary classifiers, since this is the evaluation metric at the Kaggle - Udacity challengue, this will be the metric to evaluated the selected model.



Target Value Count

1.7 Benchmark Model

As this project is part of the Kaggle - Udacity challengue, the benchmark model would be the Kaggle scores found in the leaderborad. We can see that the top 20 submission scores (AUC) are above 0.80 . Something near to 0.80 will be awesome!.

1.8 Project Design

This project will consist in the 4 following steps :

1.

Like any data science project, there's gonna be a step of data understanding and data preparation. As state in the dataset input section, there are features that will need to pass through the following steps before working with them :

- Missing values removal and/or Imputation
- Check for any outliers
- Data transformation for categorical feature in One Hot Encoding representation
- Check for correlations between the features.

2. **Customer Segmentation Report** Unsupervised learning methods will be used to analyze attributes of established customers and the general population in order to create customer

segments, identifying the parts of the population that best describe the core customer base of the company. As a baseline I will use K-means since perhaps it is the most used and well known algorith for this kind of problems. The segmentation will use the elbow method heuristic and silhouette score to find the optimal number of clusters

3. **Supervised Learning Model** Using the previous analysis, the following step will be to build a machine learning model that predicts whether or not an individual will respond to the campaign. As a baseline a will use a LogisticRegression to set up a score with a simple model. Later it will be nice to try a RandomForestClassifier or an XGBoost since they have proven to present excellent results in almost any binary clasification problem. In the creation for the supervised model, it might be a good idea to try some features scaleing techniques to ease the model fitting due to the data scales.

   The selected model will go through an Hyperparameter tuning step. Alongside with a Cross-Validation to check model performance and avoid any suspicious behavior

It's an imbalanced problem ! . There are a few options for this. In this project at lest I will try SMOTE as found it is a good technique in this kind of problems.

Libraries to use:

- NumPy——&——Pandas————> Data Manupulation.
- SageMaker–&—Scikit-Learn——> Machine Learning and Feature tranformation
- Matplotlib—&—Seaborn————> Data Visualization.

4. **Kaggle Submission** : An additional minor step to try and test the resulting model in the Kaggle Competition. Time to make some predictions and check our selected model.

1.9 References

1. Average Advertisin Costs
2. Silhoutte Score
3. Elbow Method Heuristic
4. ROC Curve AUC
5. K-MEANS