# **Scrutinizing Customer Segmentation for Arvato Financial Services**

A blog for Udacity Data Scientist Capstone Project

Problem statement: Create a Customer Segmentation Report for Arvato Financial Services.

# 

# Introduction

*The motive of this project is to predict which individuals are most likely to convert into becoming potential customers for a mail-order sales company in Germany.*

This project is a real-life data science task provided through Udacity by partners at Bertelsmann Arvato Analytics. Here, we will analyse demographics data for customers of a mail-order sales company in Germany, comparing it against demographics information for the general population. Then, we will apply unsupervised learning to perform customer segmentation, identifying the parts of the population that best describe the core customer base of the company. Finally, we will analyzy a third dataset with demographics information for targets of a marketing campaign for the company, and use a ML model to predict which individuals are most likely to convert into becoming potential customers for the company.

The final prediction can be evaluated by the means of [Kaggle competition](https://www.kaggle.com/c/udacity-arvato-identify-customers#description) with evaluation metric of [AUC for the ROC curve](https://en.wikipedia.org/wiki/Receiver_operating_characteristic#Area_under_the_curve), where the ROC curve is created by plotting the [true positive rate](https://en.wikipedia.org/wiki/True_positive_rate) (TPR) against the [false positive rate](https://en.wikipedia.org/wiki/False_positive_rate) (FPR) at numerous threshold settings. This evaluation metric is preferred because the training data about which we will talk later is highly unbalanced.

**Data Investigating and Data Cleaning**

There are 4 datasets (2 description files) associated with this project:

***AZDIAS:****Demographics data for the overall general population of Germany.*

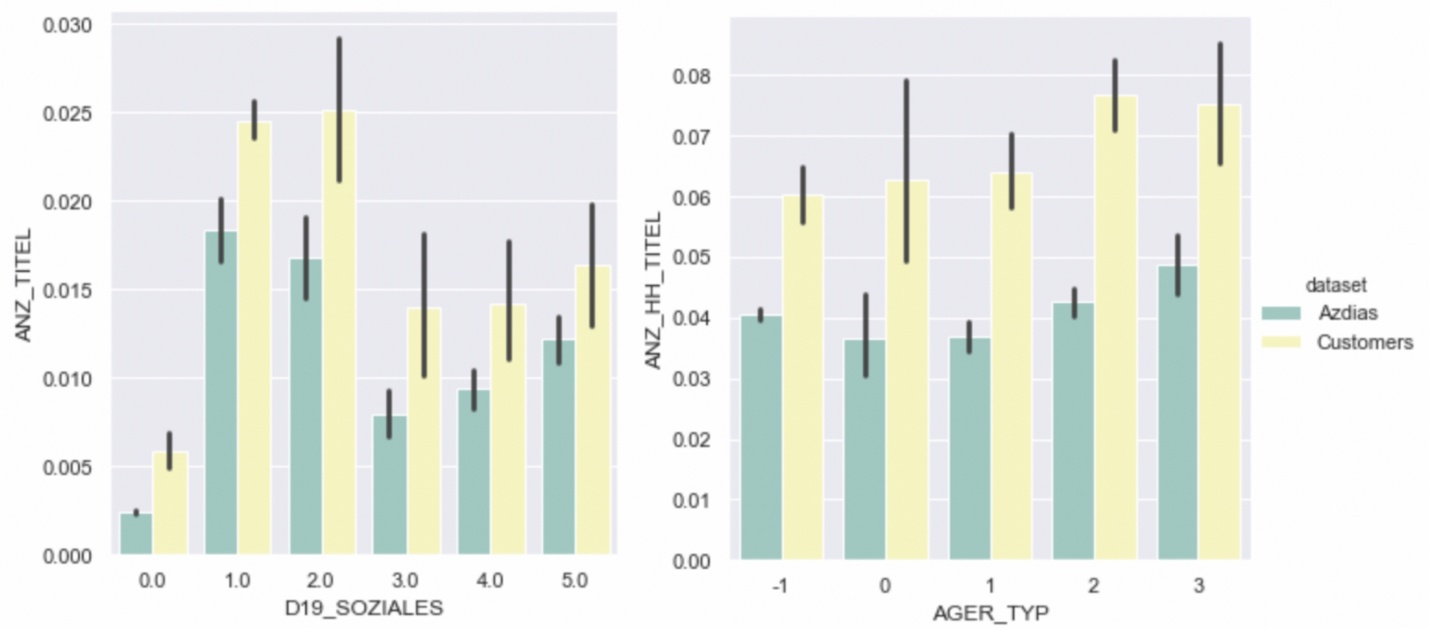
***CUSTOMERS:****Demographics data for studying the customers of a mail-order company.*

***MAILOUT\_TRAIN****and****MAILOUT\_TEST****: Demographics data for individuals who were the targets of a marketing campaign.*

We are going to compare **CUSTOMERS** opposite to **AZDIAS**and use unsupervised learning techniques to perform customer segmentation. While **MAILOUT\_TRAIN** and **MAILOUT\_TEST** will be used in the Supervised Learning part of this project, the [Kaggle Competition](https://www.kaggle.com/c/udacity-arvato-identify-customers) is to identify the parts of the population that best describe the core customer base of the company.

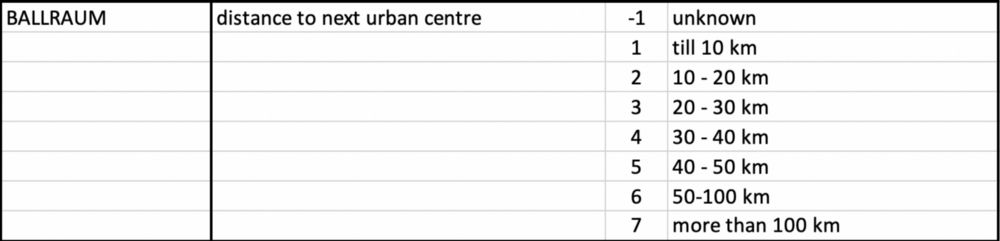
Glance of AZDIAS dataset

AZDIAS dataset consistes of 891211 persons (rows) x 366 features (columns). The dataset is majorly composed of ordinal and categorical data which were label encoded and few of them have high percentage of missing value.

Firstly, let’s take a look at the distribution and comparison of some random features between AZDIA and CUSTOMERS dataset.

Random feature distribution difference between Azdias and Customers dataset

Here I randomly choose two categorical data and ploted them against two random numerical data to check if there is any difference between the distribution of Azdias and Customers dataset. After having some general knowledge of two main dataset, we can start with some feature engineering and data cleaning.



Description files gives a detailed information of the meaning of each label including how missing value are labeled. We need to map the label of missing value back to NaN and analyse the number of missing value in each columns. Lastly, columns with missing value higher than 20%.

Since the unsupervised learning techniques used we will only work on data that is encoded numerically, we need to make a few encoding changes or additional assumptions to be able to make progress.

Mixed Type Column: PRAEGENDE\_JUGENDJAHRE

For example, we need to notice the column with mixed type interpretation. These columns combines information on 3 dimensions: generation by decade, movement (mainstream vs. avantgarde), and nation (east vs. west). We have to manually extract the information and generate new columns to store the data.

Summarising the steps:

*Assign unknown data back to NaN and drop the columns with higher percentage of missing value.*

*Investigate mixed type columns and extracted useful information to form new columns.*

*Apply one-hot encoding to categorical data.*

*Study and pre-process columns with time-related features.*

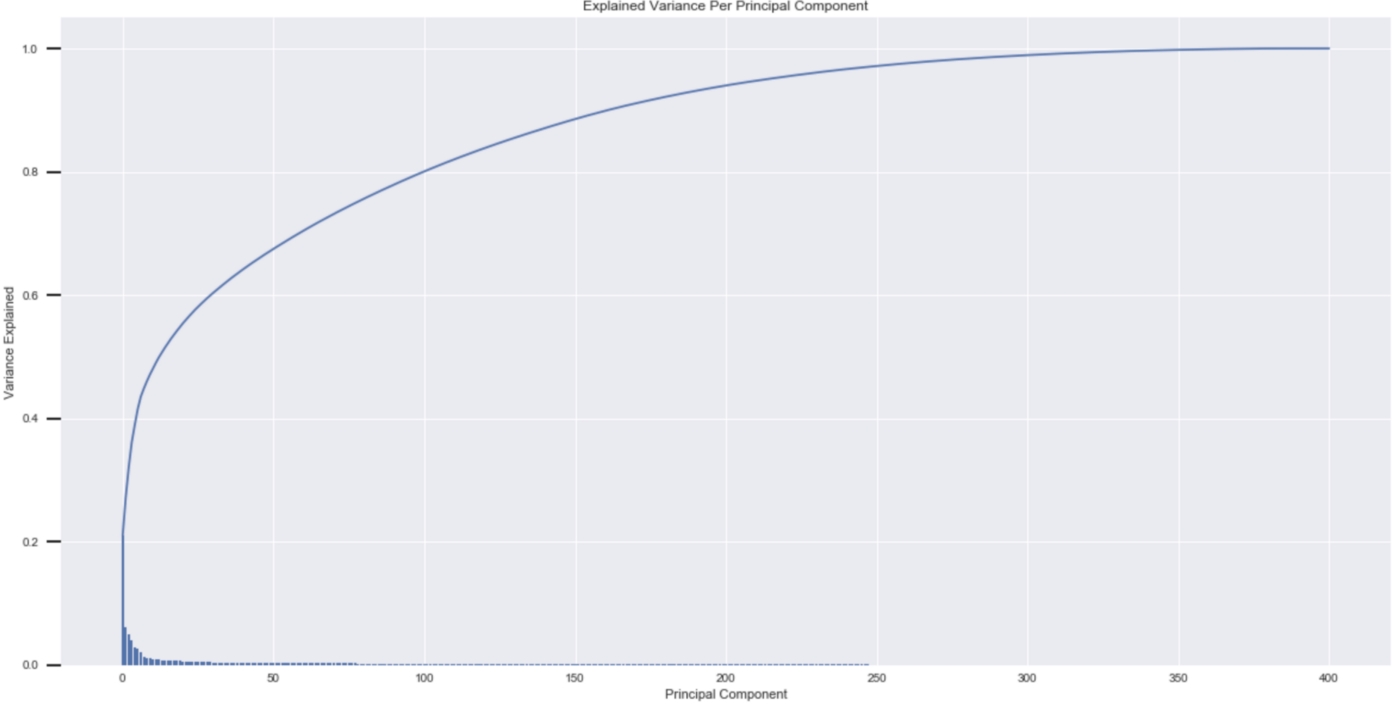
*Apply standard scalar to particular columns and fill missing value with ‘mean’.*

# Customer Segmentation

In this section, We will go through how I utilized unsupervised learning to analyse attributes of established customers and the general population in order to create customer segments.

## 1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is one of the most essential techniques in Exploratory Data Analysis to understand the data, reduce its dimensions and for entire unsupervised learning in general. Based on the cumulative variance, we will decide the number of transformed features to be retained.

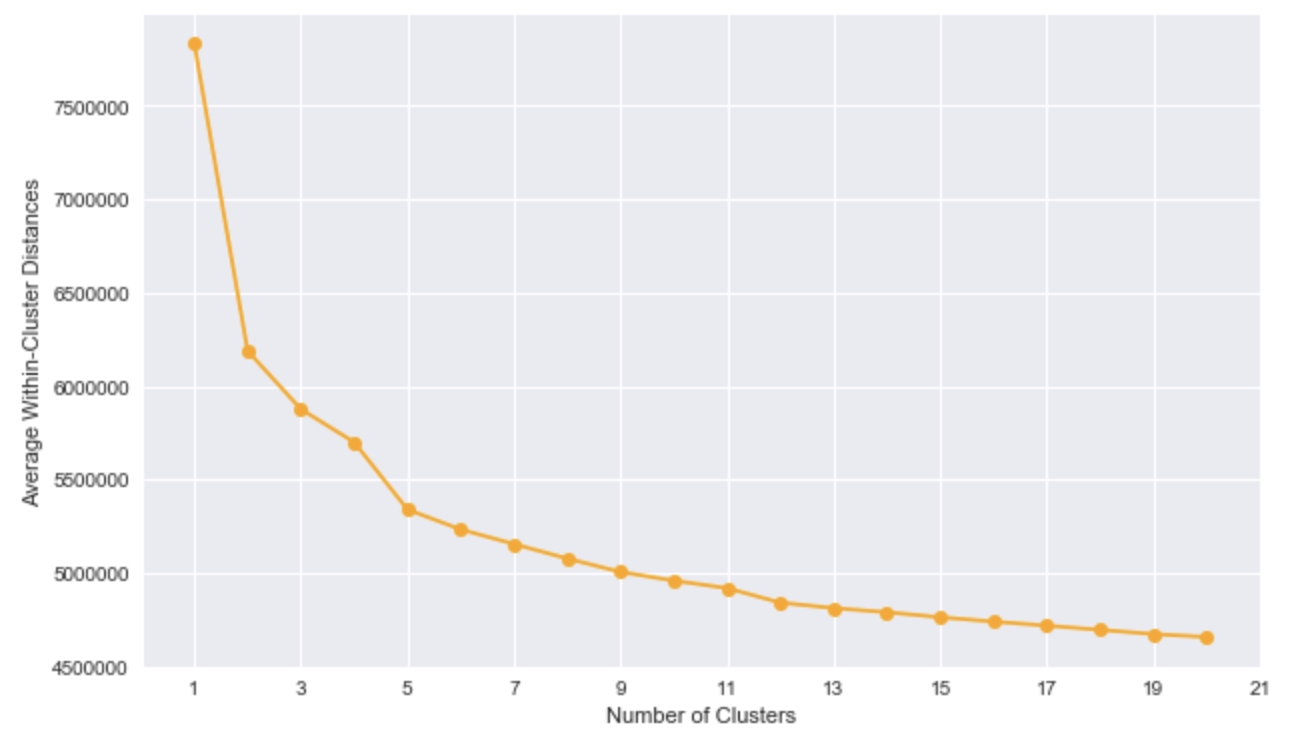


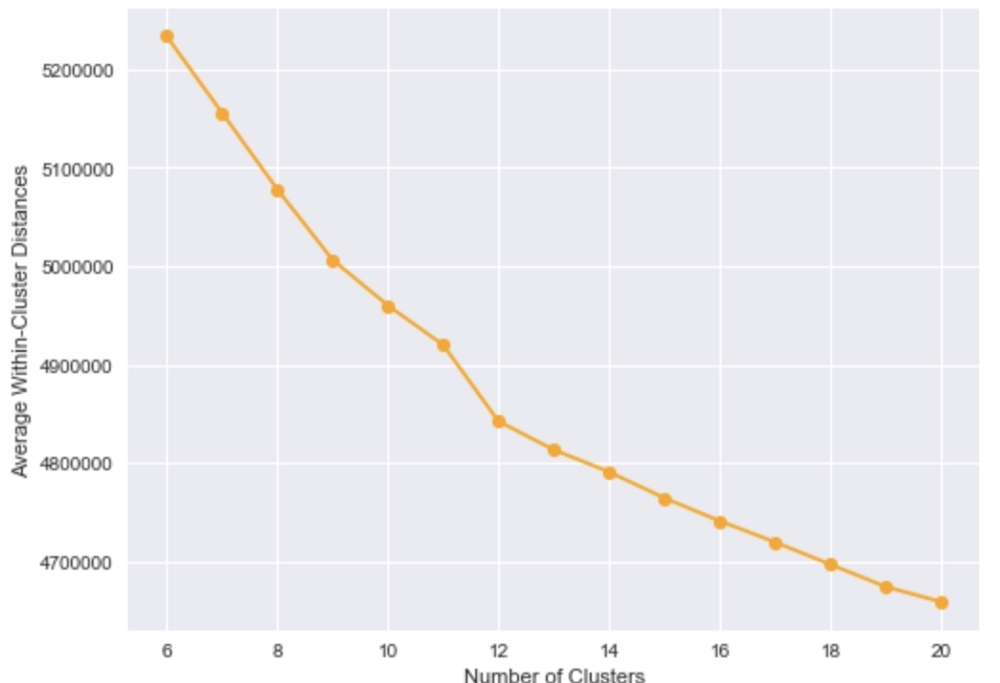
PCA ratio of variance and cumulative variance

At the end, I kept 260 principal components and the cumulative variance of more than 95%. We can always increase the number and retrain the PCA model if the future performance is not satisfied.

## 2 Cluster

After the number of principal components to be retained is decided, the next step is to see how the data clusters in the principal components space. Here we will apply k-means clustering to the dataset and use the mean within-cluster distances from each point to their assigned cluster’s centroid to decide the number of clusters to keep.

Average Distances vs Number of Clusters

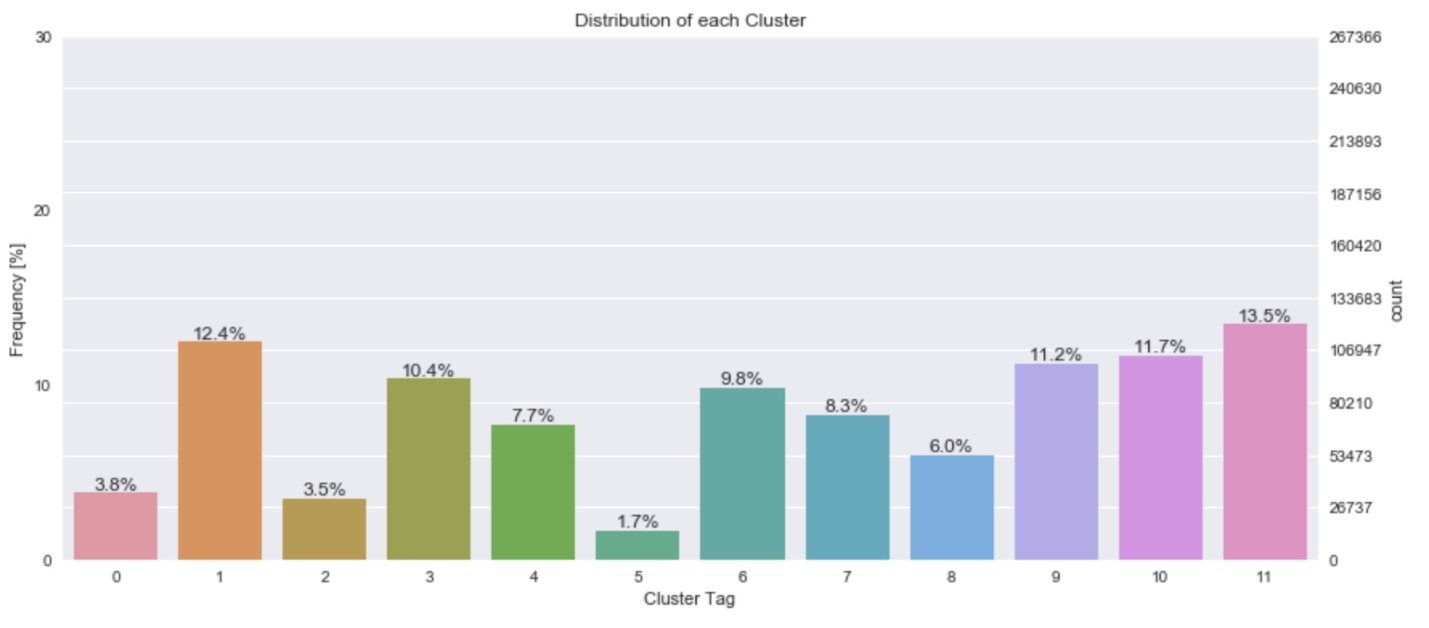


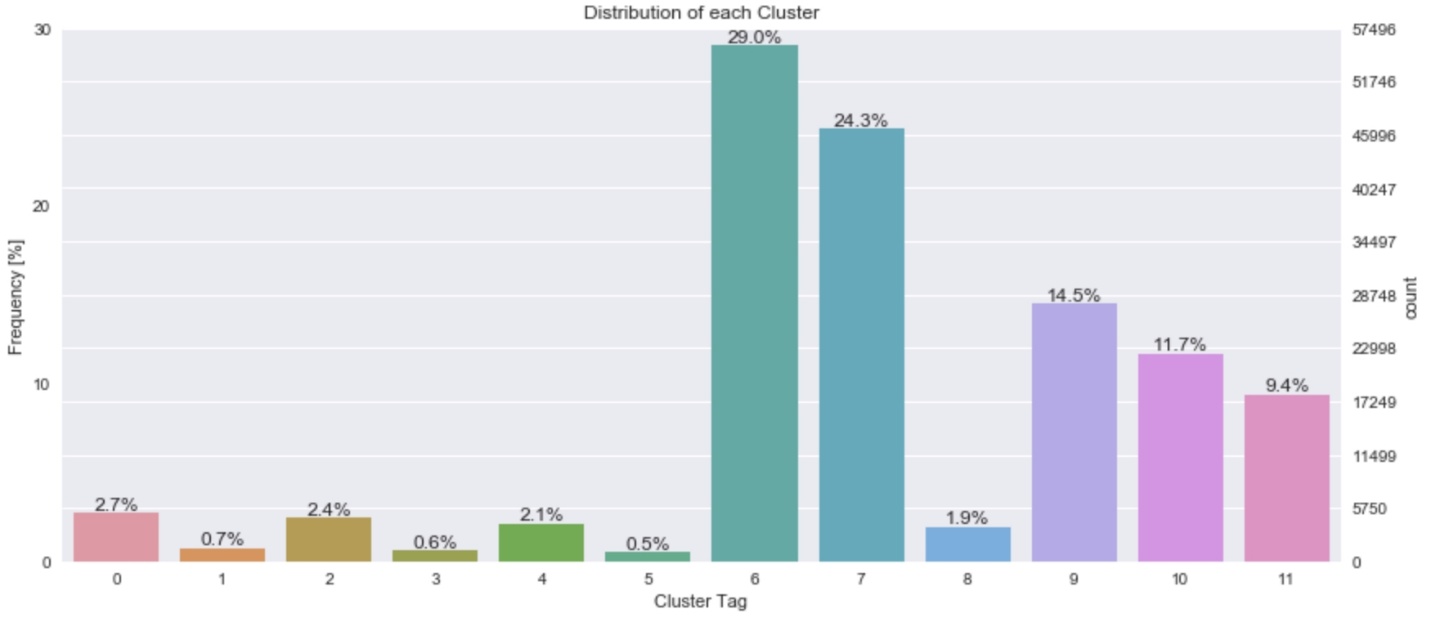
Zoom-in graph of Distances vs Number of Clusters

Usually we look at the turning point or elbow where the error starts to decrease slowly. In out graph, the distances decreases gradually after number of clusters reaches 12, as there is a relatively larger drop in distance at 12. As a result, we choose 12 as our final cluster number and the next step is to re-fit a K-Means instance to perform the clustering operation and apply it to our **CUSTOMERS** dataset.

## **3 Compare Customer Data to Demographics Data**

In this section , we are going to cluster the data based on demographics of the general population of Germany, and see how the customer data for a mail-order sales company maps onto those demographic clusters. And lastly, we will compare the two cluster distributions to observe where the strongest customer base for the company lies.

Cluster Distribution of the General Population of Germany



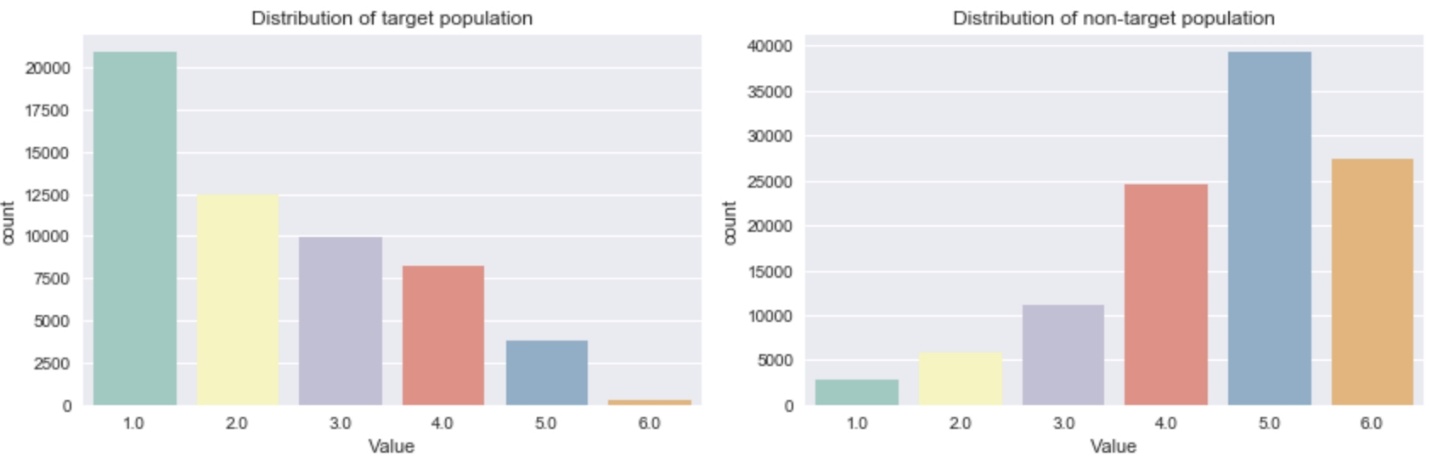
Cluster Distribution of Customers of mail-order sales company

Here we have two graphs representing the distribution of general population and customers of the company respectively. We can see that the frequencies of cluster 6 and 7 in **CUSTOMERS** data set are remarkably larger than the frequencies in **AZDIAS**(or general) dataset.

Consider the proportion of people in each cluster for the general population, and the proportions for the customers. If the company’s customer base be universal, then the cluster assignment proportions should be fairly similar between the two. The mismatch of clusters between **AZIDAS** and **CUSTOMERS** dataset shows that there are only certain segments of the population that are interested in the company’s products.

The huge difference of cluster 6/7 (29% vs 9.8% and 24.3% vs 8.3%) between two datasets suggest that the people in these two clusters are the target audience for the company. On the other hand, the proportion of the data in a cluster being larger in the general population than the customer data (e.g. Cluster 1, 3 and 8) suggests that group of people to be outside of the target demographics.

In order to get deeper insights of the cluster information, we are going to investigate Cluster 6 vs Cluster 1 to see if there is any significant difference in some specific data as well.



Comparison of column ‘**HH\_EINKOMMEN\_SCORE**’ between Cluster 6 and 1

Let’s take a look at data ‘**HH\_EINKOMMEN\_SCORE**’ which represents the **estimated household net income.**Here we can observe that there is a relatively significant difference where the targeted population (Customers) has lower value in this columns than non-targeted pupulation (General Population).

Here is the description of the data in provided excel file sheet:

*1, 0: unknown*

*1: highest income*

*2: very high income*

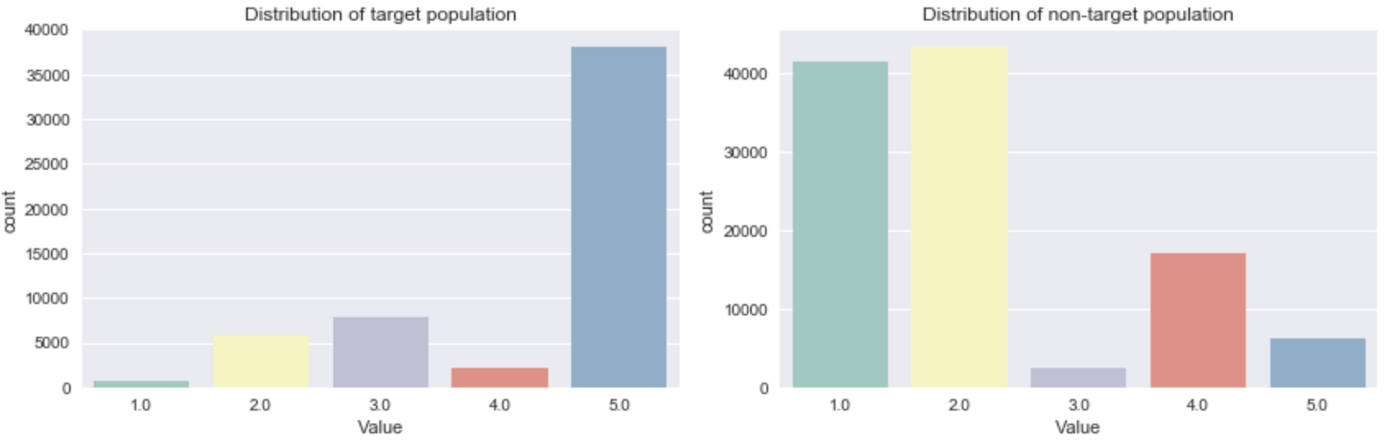
*3: high income*

*4: average income*

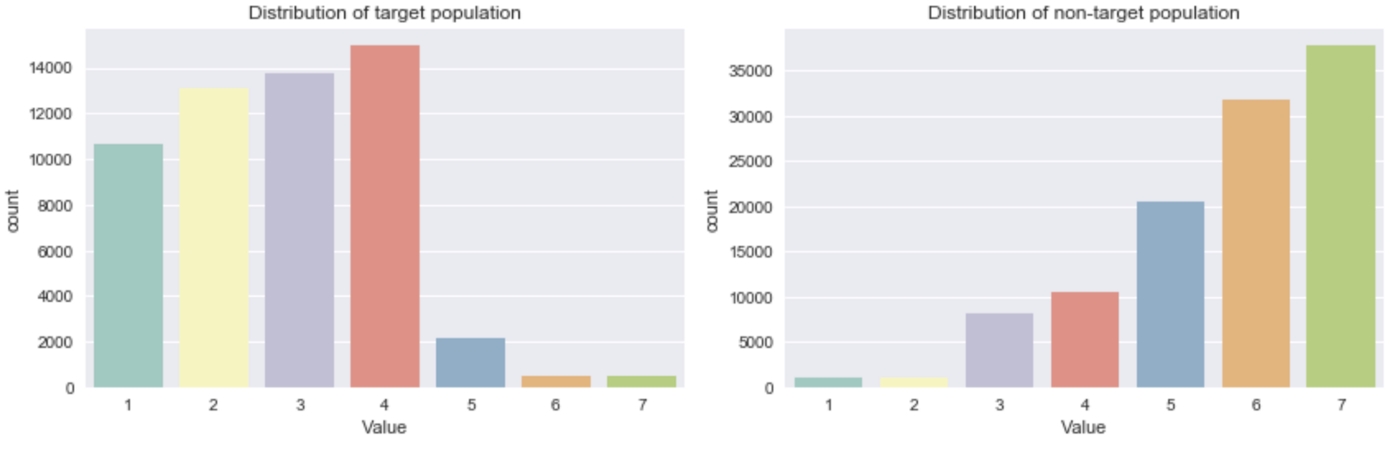
*5: lower income*

*6: very low income*

Obviously, the company should target wealthier people than people with lower income. PCA/KMeans algorithm successfully capture the differences.



Comparison of column ‘**LP\_STATUS\_GROB**’ between Cluster 6 and 1

Comparison of column ‘**SEMIO\_PFLICHT**’ between Cluster 6 and 1

I then randomly picked some columns to compare the differences between Cluster 6 and Cluster 1. Here ‘**SEMIO\_PFLICHT’**represents **Personality typology\_dutiful**where lower value means higher affinity. There are some other personality related features in the dataset that were distinguished by the algorithm and are utilized to help target the customers.

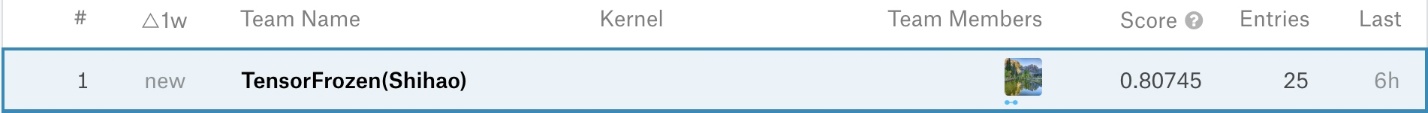
# Supervised Learning and Kaggle Competition

Lastly, in the last part of the project, we are going to apply supervised learning to investigate **MAILOUT\_TRAIN**and**MAILOUT\_TEST**datasetto predict whether or not a person became a customer of the company following the campaign. (Represented by ‘**RESPONSE**’ column in **TRAIN** dataset.)

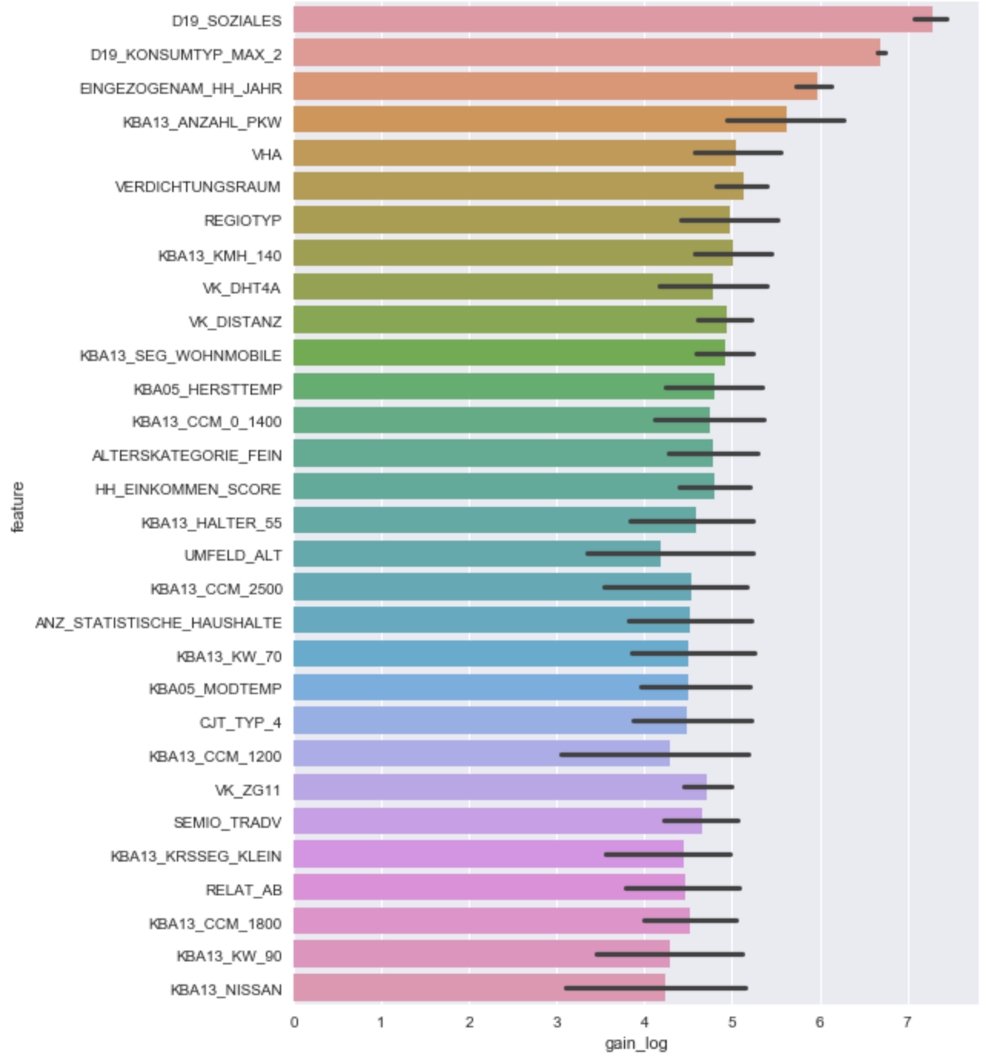
Here the **MAILOUT** dataset has been split into two approximately equal parts, each with almost 43 000 data rows. We are going to predict the ‘**RESPNESE’**for **MAILOUT\_TEST**dataset and the prediction can be submit to [**Kaggle**](https://www.kaggle.com/c/udacity-arvato-identify-customers) for evaluation.

After some quick investigation against **MAILOUT\_TRAIN**dataset, we can find over 43 000 individuals, only 532 people response to the mail-out campaign, which means the training data is highly unbalanced. Based on this discovery, we have to split the data using **StratifiedKFold** based on the distribution of target.

There are many machine learning models we can choose, starting from linear model such as ‘**LogisticRegression’**ortree based model such as ‘**DecisionTreeRegressor’** toensemble models such as **‘RandomForestRegressor’**and **‘GradientBoostingClassifier’**. After testing the performance of several models based on cross-validation result, I used **‘LGBMRegressor’**, andtrained the model by using a 5 folds validation method. At the end, I submit the result to Kaggle competition and achieved a 0.80745 **roc\_auc\_score**(Receiver Operating Characteristic Curve).



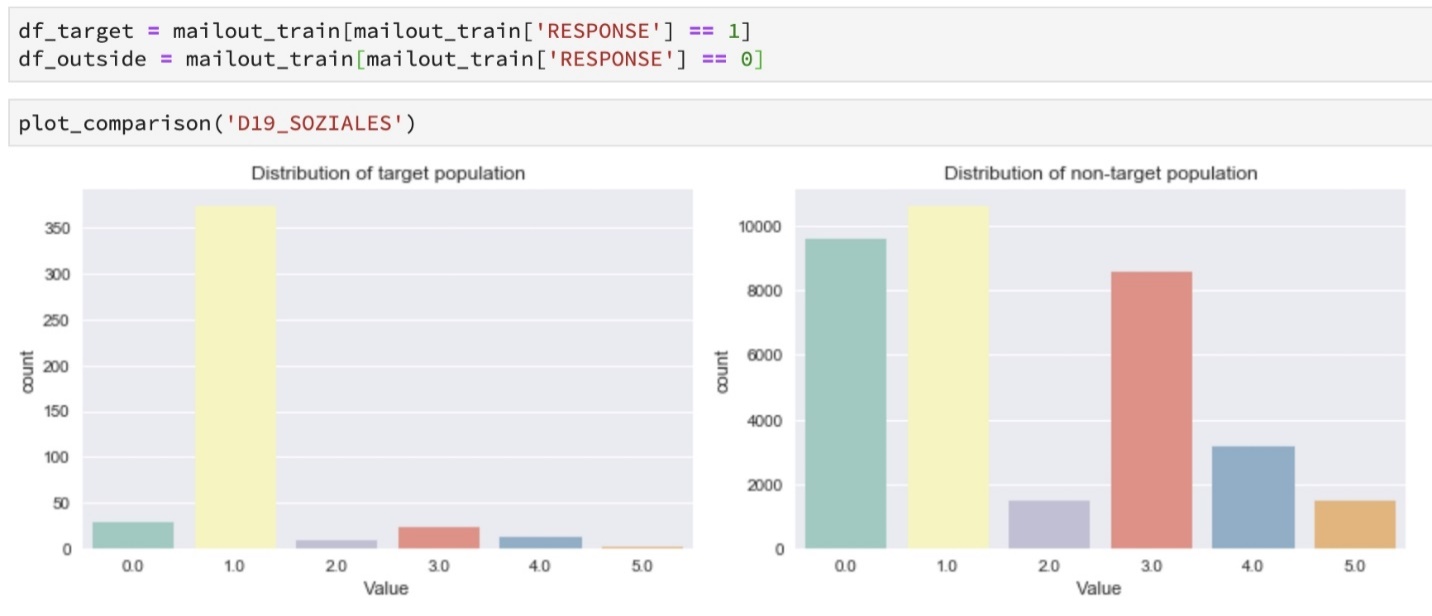
There are a lot of different approaches that can be applied from data cleaning to model training.

Feature Importance of LGBMRegressor

A lot has been said, now let’s take a look at the result of the **LGBMRegressor.**

The left graph shows the feature importance of the LGBMRegressor result. The x value is calculated by mean\_log gain value to illustrate the importance of features.

Let’s take a look at the most important feature ‘**D19\_SOZIALES**’.

Comparison of feature ‘**D19\_SOZIALES**’ between Response/Non-Response Population

From the upper graph, we can conclude that almost all the Response individuals lies in the value 1 in feature ‘**D19\_SOZIALES’**. Although there is no description provided for this feature, features starting with **D19**are mostly related to transactional activities based on certain product group.

# Conclusion

In this blog, we dive into a real life machine learning project provided by Arvato Financial Solutions, a Bertelsmann subsidiary.

*We Investigated Demographics data of general population of Germany and data for customers of a mail-order company.*

*Pre-processed the dataset based on column/feature property.*

*Applied Unsupervised Learning Algorithms, namely PCA and KMeans to segment the population (into different clusters) to recommend the potential customers for the company.*

*We took a deeper look at two main clusters and compare them by checking the differences of several randomly choose features.*

*Apply Supervised Learning to predict whether or not a person became a customer of the company following the campaign.*

*Investigated the most important feature trained by machine learning model and compare the feature distribution between target/non-target population.*