Arvato Identify Customer Segmentation Final Report

Author : Hafizur Rahman

Overview

This project is related to demographics data for customers of a mail-order sales company in Germany. We aim to use both unsupervised learning and supervised learning to tell the difference between the general population and mail-order customers. In this report, I followed the traditional steps of data analyzing work-flow, and cover the following contents:

- Data description
- Data Visualization
- Data cleaning
- Unsupervised learning
- Supervised learning
- Conclusion

Data description

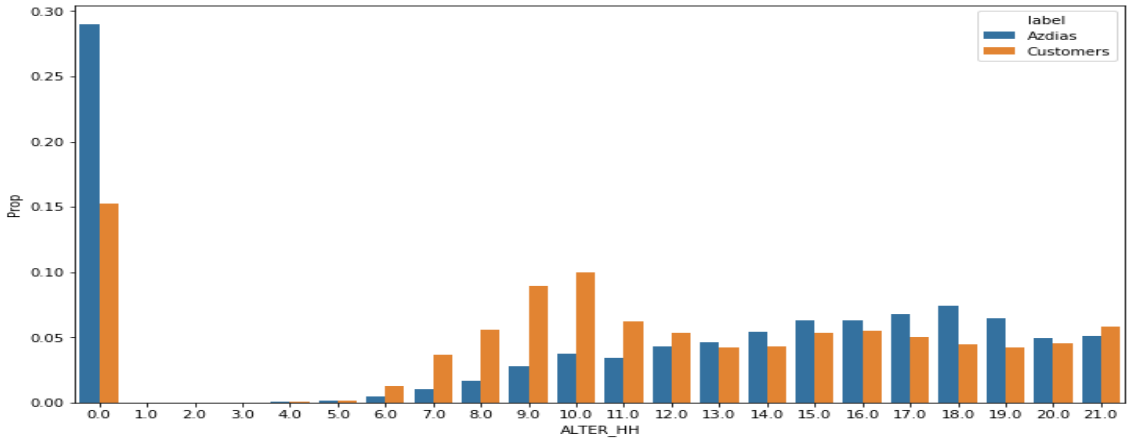
There are four data files in this project:

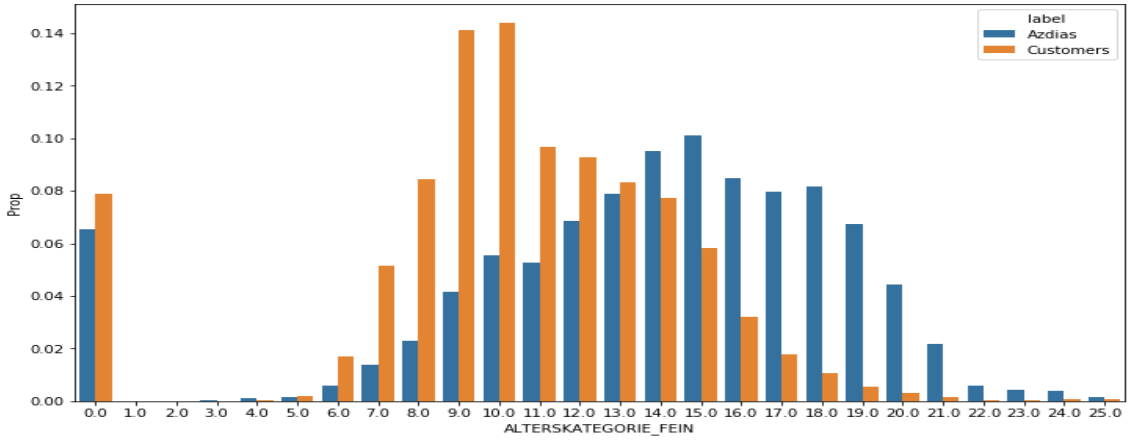
- Udacity_AZDIAS_052018.csv: Demographics data for the general population of Germany; 891 211 persons (rows) x 366 features (columns).
- Udacity_CUSTOMERS_052018.csv: Demographics data for customers of a mail-order company; 191 652 persons (rows) x 369 features (columns). Comparing with “AZDIAS”, The “CUSTOMERS” file contains three extra columns (‘CUSTOMER_GROUP’, ‘ONLINE_PURCHASE’, and ‘PRODUCT_GROUP’). We will use the first two datasets to do unsupervised learning and try to describe which parts of the general population that are more likely to be part of the mail-order company’s main customer base, and which parts of the general population are less so.
- Udacity_MAILOUT_052018_TRAIN.csv: Demographics data for individuals who were targets of a marketing campaign; 42 982 persons (rows) x 367 (columns). We will use this dataset to train our supervised learning model.
- Udacity_MAILOUT_052018_TEST.csv: Demographics data for individuals who were targets of a marketing campaign; 42 833 persons (rows) x 366 (columns).

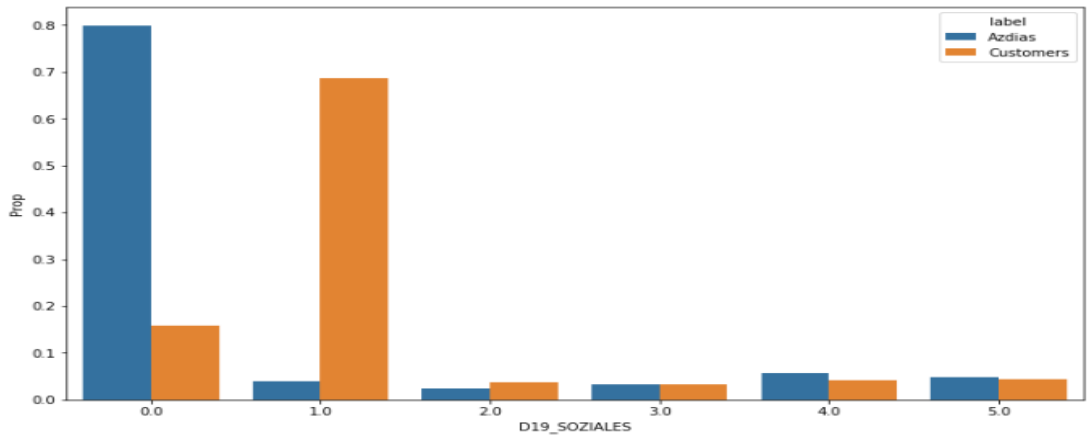
Data Visualization

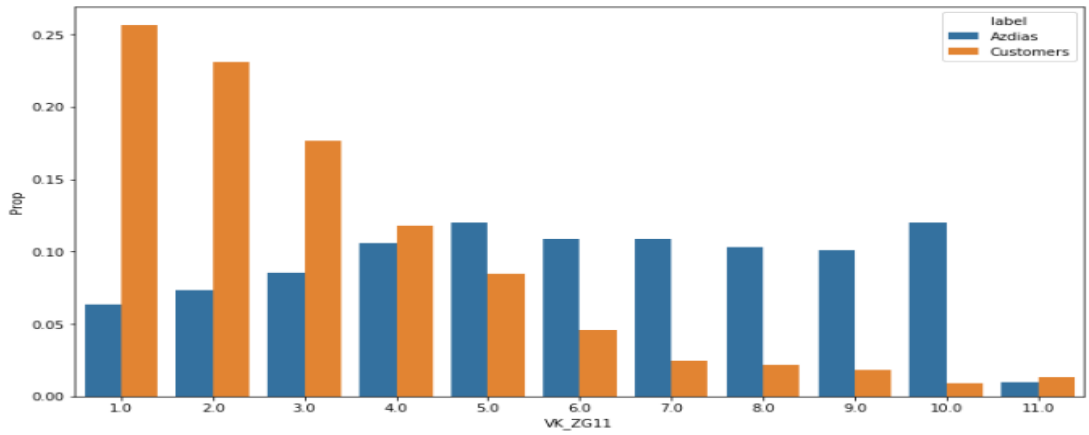
Data visualization is an excellent way to understand the data. In this section, I plotted out all the features' distribution in both AZDIA and CUSTOMERS dataset. Some features have similar distributions, while some have obviously different distributions. Here I chose some of them to display.

Features have similar, also different distributions:


label
Azdias
Customers
Prop
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0 9.0 10.0 11.0 12.0 13.0 14.0 15.0 16.0 17.0 18.0 19.0 20.0 21.0
ALTER_HH


label
Azdias
Customers
Prop
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0 8.0 9.0 10.0 11.0 12.0 13.0 14.0 15.0 16.0 17.0 18.0 19.0 20.0 21.0 22.0 23.0 24.0 25.0
ALTERSKATEGORIE_FEIN





Data cleaning

After the data visualization, we already have some basic understanding of the dataset; then we could do the data cleaning. To simplify this step, I merged the “AZDIAS” and “CUSTOMERS” and added an extra column to label which dataset the data is from.

The new dataset is 1082873 persons (rows) x 370 (columns). For the columns, nine of them are “Object” types.

```
shape after corr (733227, 238)
shape after one-hot (733227, 284)
shape after impute (733227, 284)
inside outliers if
shape before scaling (415405, 284)
shape after scaling (415405, 284)
(415405, 283)
```

```
shape after corr (191652, 256)
shape after one-hot (191652, 303)
shape after impute (191652, 303)
inside outliers if
shape before scaling (100341, 303)
shape after scaling (100341, 303)
(100341, 302)
```

- Column ‘CAMEO_DEUG_2015’, ‘CAMEO_DEU_2015’ and ‘CAMEO_INTL_2015’ have some special characters “X” and “XX” in them.
- Some numbers in column ‘CAMEO_DEUG_2015’ and ‘CAMEO_INTL_2015’ are “string” format. — Features of “CAMEO_DEU_2015”,”CUSTOMER_GROUP”,”D19_LETZTER_KAUF_BRANCHE” and “OST_WEST_KZ” are categorical variable. It is necessary to encode them first.
- The column “EINGEFUEGT_AM” is a data time format. I extracted the year from the time and deleted the original column.
- The column “CUSTOMER_GROUP” and “PRODUCT_GROUP” are only contained in the “CUSTOMERS” dataset. They are dropped later.

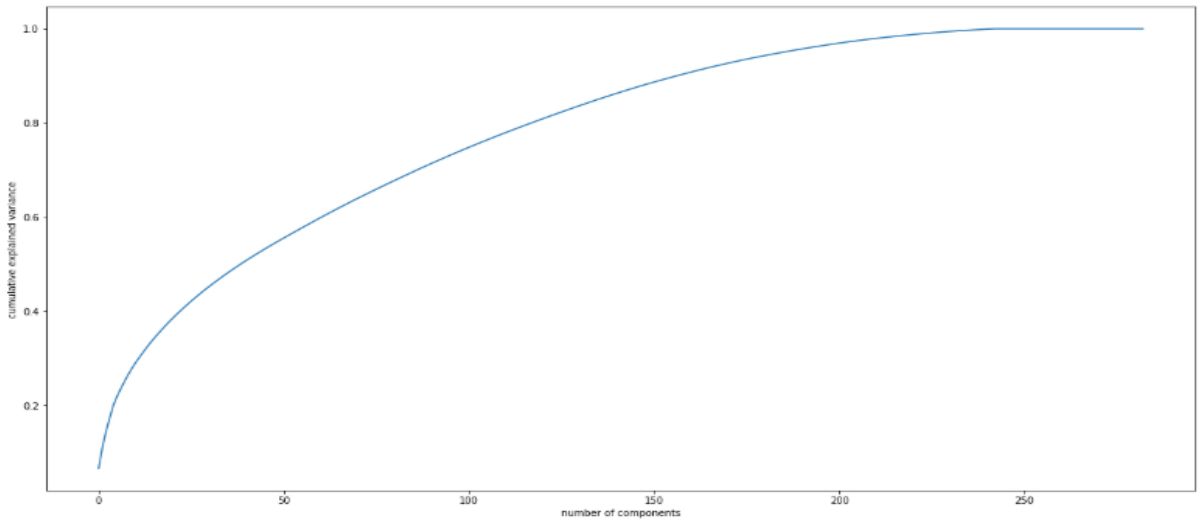
The next step is to deal with the missing values. According to the description file, some data is labeled as -1 and 0, when this data is unknown. It belongs to the missing values. We should change those data into Nan first before investigating the number of missing values in each column. Here is the columns list which has missing values higher than 20%. I dropped these columns before the unsupervised learning and filled the missing values with -1 for the columns with few missing values.

Unsupervised learning

In this part, we will use the K-means clustering method to perform unsupervised learning. While because the data size is enormous, firstly we try to reduce the dimensions of the dataset. The most common dimensions reduction method is PCA (Principal Component Analysis). It could speed up the computation in the case of high dimension dataset.

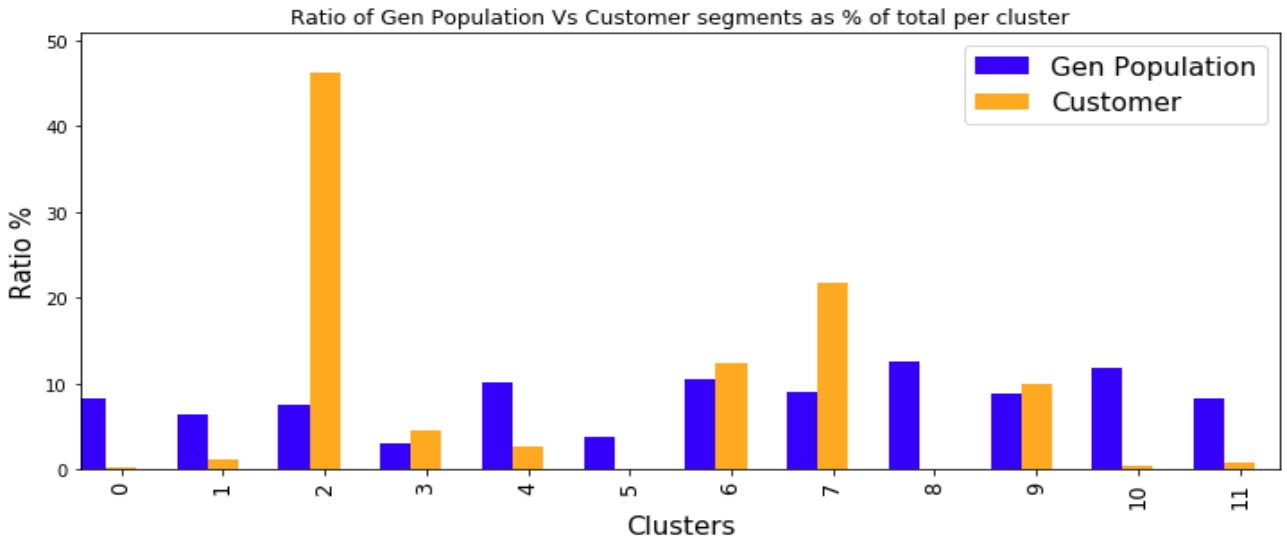
- **PCA Visualization**

To prepare the data for unsupervised learning, we first use PCA to reduce the dimension. The figure below gives out the change of cumulative variance with the number of principal components. From the graph, we can see when the number of principal components is 300; the cumulative variance is more than 98%.



- **K-Means**

After PCA, I applied K-means clustering to perform unsupervised learning. For K-means clustering, we first need to decide the number of clusters to split. The figure below shows the sum of every points' distance to its cluster center. We can see when the cluster number is 12, the slope of the line turns to small. Thus we choose eight as the number of clusters.



The next figure shows the distribution of customers and the general population in each cluster. It is interesting to find that most individuals in cluster 2 are the general population and mail-order customers have a large portion of people in cluster 6 and 7.

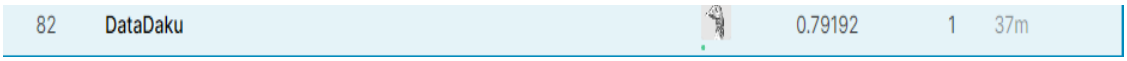
Supervised learning

In this part, I used the supervised learning method to predict whether or not a person will become a customer based the demographic information.

There are two datasets in this part, MAILOUT_TRAIN, and MAILOUT_TEST. MAILOUT_TRAIN includes a column “RESPONSE,” that states whether or not a person became a customer of the company following the campaign. I will use this dataset to training my model and then create predictions on the “MAILOUT_TEST” dataset. We will apply the AUC metrics to evaluate the model ’s performance.

There are many supervised learning methods which could be used in this case, such as “Decision tree”,”Random forest”,”SVM”,”LGBM”. At the

very beginning, I used the “XGBoost”, but I only get a 0.5000 roc-auc-score. After having a deeper insight into the training data, we find it is highly biased data. In the training dataset, there are 42962 individuals, but only 532 of them response to the mailout campaign. Then I tried the “LGBMRegressor” which achieved a much better result than before. In the Kaggle competition, my final score is 0.79192 (The leading score is 0.81063).



Conclusion

In this work, we first explored the data, and then we built both unsupervised and supervised learning method for analyzing the data from Arvato Financial Solutions.

By the unsupervised learning k-means, we compared some features which may relate to whether or not a person will become a customer. In the supervised learning GradientBoostingRegressor, we made a prediction based on the demographic information and achieved a good roc-auc-score. In real business, the company could send invitations according to the predicted ranking list.