Machine learning Nanodegree

Capstone project report

Bertelsmann/Arvato customer segmentation report

Ahmed Magdy Ali

December 20, 2020

Definition

Project overview:

Arvato is a global services company headquartered in Germany Its services include customer support, information technology, logistics, finance, and aftersales solutions. In this project I will use the dataset provided by Arvato to perform customer segmentation based on demographic features of customers of the company using unsupervised learning, also predict individuals that are most likely to be potential customers of the company using supervised learning.

Problem Statement:

The problem is: given demographic features of population and features of customers, how likely an individual be a customer for the company.

Solutions: understanding the given data, looking for missing values of features and handling it then dealing with categorical values, then discovering the demographic features for the customers of the company by studying features of population and customers of the company using unsupervised learning techniques, then predicting if a person is most likely to be a customer for the company or not by using supervised learning method.

Evaluation Metrics:

The evaluation metric on kaggle is AUC-ROC or “Area Under the Curve” (AUC) of “Receiver Characteristic Operator” (ROC) which is evaluation metric for binary classification problems. It is a probability curve that plots the TPR against FPR at various threshold values. The AUC performance metric is appropriate for the imbalanced labeled dataset as the number of negative labels is significantly higher than the number of positive labels.

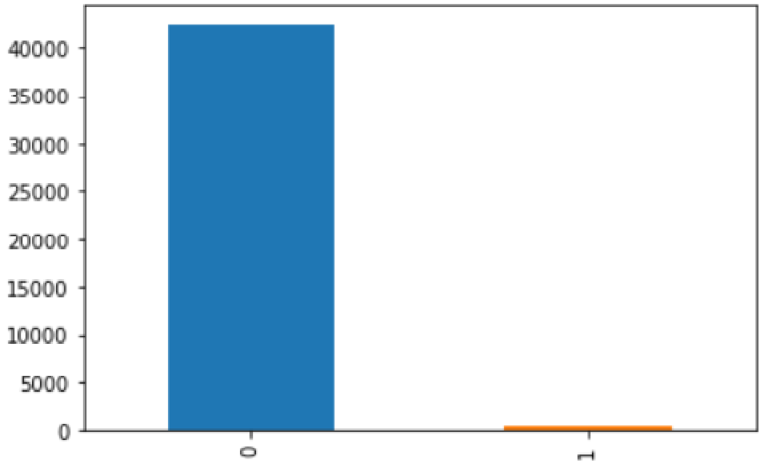


Figure 1: Labels of Test Data Distribution

Analysis

Data Exploration and Preprocessing:

Workflow:

Exploring the data and preprocessing by looking for missing values and handle it, also any categorical data should be encoded and converted into numerical data, then scaling the data to make sure the values are in the same domain. After that decide what are the necessary features to keep and what are the unnecessary features to drop, I have implemented functions to do all the preceding steps to facilitate the use of them since there are a lot of features in our dataset, we need a dimensionality reduction technique like PCA Using unsupervised learning algorithm to segment the data of population and customers into different segments. After that preprocessing the train and test data then by using a supervised learning method train the model and then make predictions on the test data.

Detailed exploration and preprocessing:

Data exploration:

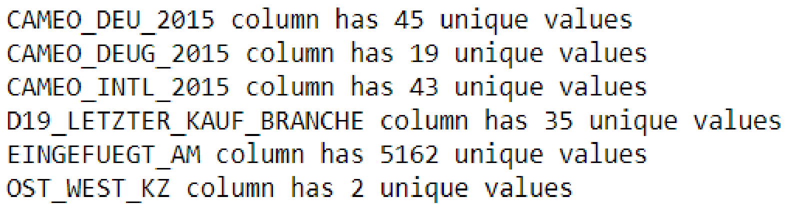
first I created explore function to print the shape, first few rows and a description for the dataset. After that I checked the warning about the 18th and 19th columns that it contains mixed datatypes, I found that is contains 'X' and 'XX' values within it so i checked the DIAS Attributes file I found that -1 indicates that this value is unknown value so I decided to change it to -1 as the description, also I found that some of the inputs where float values and others were integers inside string quotes so I decided to change the datatype of these two columns into float.

Handling category values:

I checked for the category columns I found 6 columns which are 'CAMEO_DEU_2015','CAMEO_DEUG_2015', 'CAMEO_INTL_2015',

'D19_LETZTER_KAUF_BRANCHE', 'EINGEFUEGT_AM', 'OST_WEST_KZ'

Then I realized that:



```
CAMEO_DEU_2015 column has 45 unique values
CAMEO_DEUG_2015 column has 19 unique values
CAMEO_INTL_2015 column has 43 unique values
D19_LETZTER_KAUF_BRANCHE column has 35 unique values
EINGEFUEGT_AM column has 5162 unique values
OST_WEST_KZ column has 2 unique values
```

I checked for 'D19_LETZTER_KAUF_BRANCHE' in the DIAS Attributes and DIAS Information Levels files I found no description for it and as it has many unique values I decided to drop it with 'EINGEFUEGT_AM' column.

Handling missing Values:

Column wise:

I plotted the columns with the highest percentage of missing values in descending order, after analyzing the percentage of missing values distribution I found that 6 columns (ALTER_KIND4, ALTER_KIND3 ,ALTER_KIND2, ALTER_KIND1, EXTSEL992, KK_KUNDENTYPE) has more than 30% missing values and these 6 columns has at least 65% missing values which is a huge percent so I decided to drop them, I also dropped the 'LNR' column as it was like ID so it won't benefit my model.

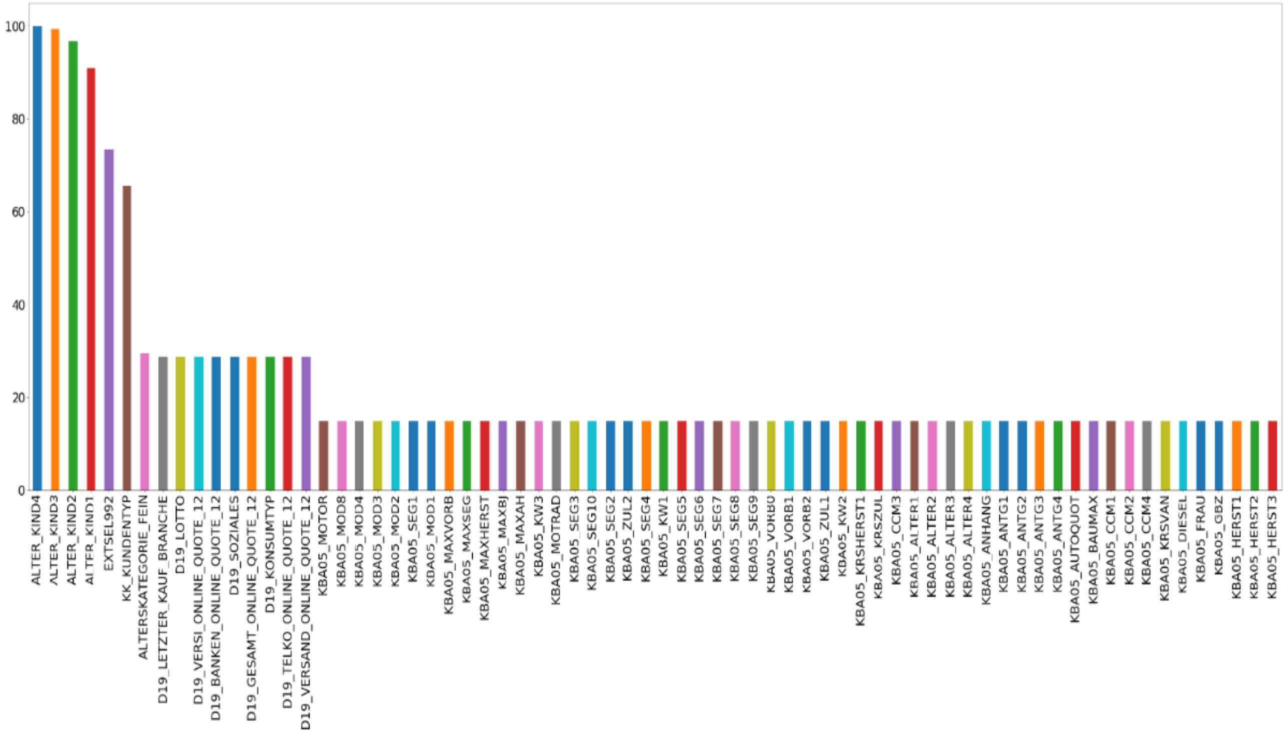


Figure 2: Top 70 Rows with missing values

Row wise:

In azdias dataset I found that there was about 82% of the data has 16 missing values or less so I decided to drop all rows that has more missing values.

In customers dataset I found that about 70% of data has 16 missing values or less and the customers dataset is way less than the population's dataset so I decided to not drop any rows in it.

Correlation:

I check for highly correlated features above threshold 0.9 and drop them to decrease number of columns a bit and also speed up learning algorithm

Imputer:

Then I used imputer with (most_frequent) technique to handle all the messing values in both datasets.

Feature Scaling and standardization:

A standard scaler is used to bring all the features to the same range. The idea behind Standard Scaler is that it will transform your data such that its distribution will have a mean value 0 and standard deviation of 1.

Manually Dropped columns and the cause of dropping:

-'D19_LETZTER_KAUF_BRANCHE': No description and too many unique values.

-'EINGEFUEGT_AM': Huge number of unique values, and it consists of dates which might not be helpful for the model.

-'LNR': Like ID column and won't benefit the learning process of the model.

-''CAMEO_DEU_2015'': So many unique values and it is like the 'CAMEO_DEUG_2015' so I decided to drop it instead of encoding it.

Algorithms and Methodology

Customer segmentation using unsupervised learning:

This part of project aims to describe the relationship between the demographics of the company's existing customers and the general population of Germany by dividing customers and the general population into segments, then compare them to be able to determine the characteristics of population to target that would be likely future customers.

Dimensionality Reduction:

Since we were left with 333 rows in the datasets after the preprocessing which is still a huge number of rows, I decided to decrease this number and at the same time discover which features will be able to explain the variance of the dataset, so I used PCA to do this and also help to ease the learning process of the model in less time, so I plotted the variance in the azdias dataset and found that only 230 component would be enough to explain 95% of the variance in the dataset.

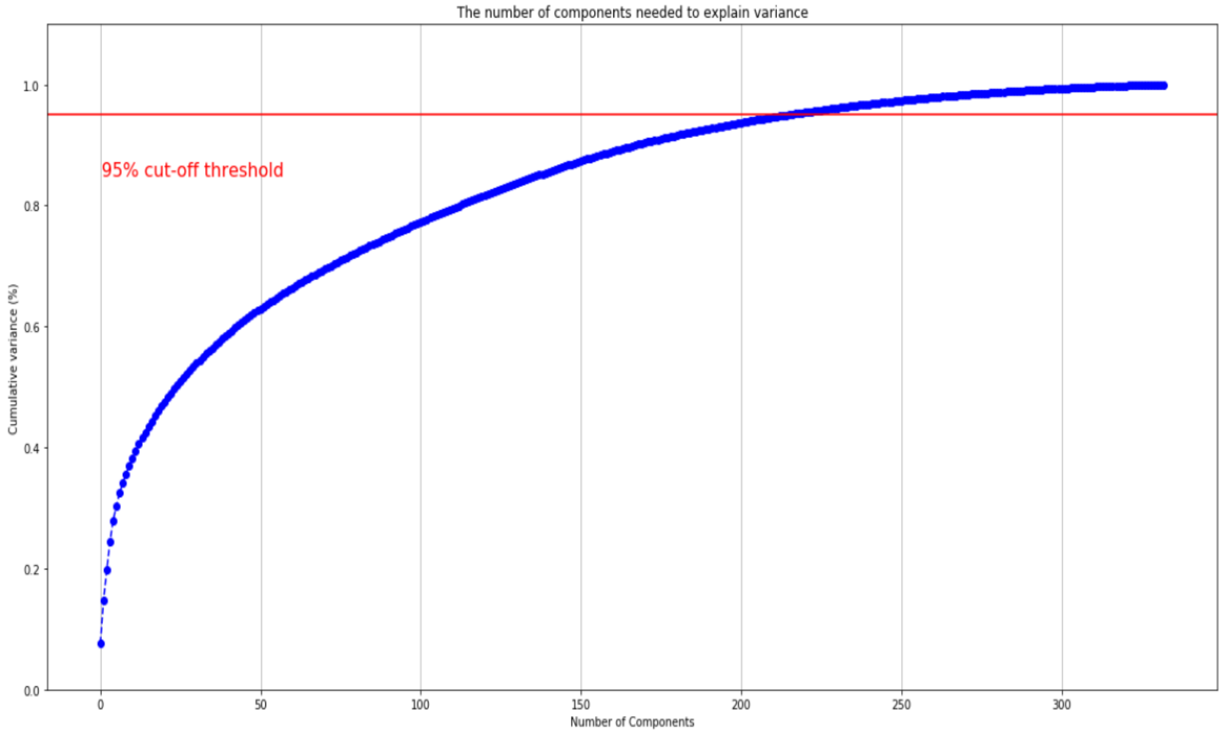


Figure 3: PCA Variance Plot

Some PCA Components Analysis:

At component 0 the people tend to have a fancy lifestyle and they have common traits like fancy cars like BMW and new cars, and their financial topology isn’t money savers or investors..

At component 1 the people have same movement routine and their social status is fine and they tend to live in 1-2 families in the same house and less than 10 families in the house.

Kmeans Clustering:

Now I have to cluster customers and general population into different clusters using Kmeans algorithm, and then use these clusters information to find similarities between general population and customers.

To decide the right number of K which is the number of clusters for the Kmeans Algorithm I used the elbow method.

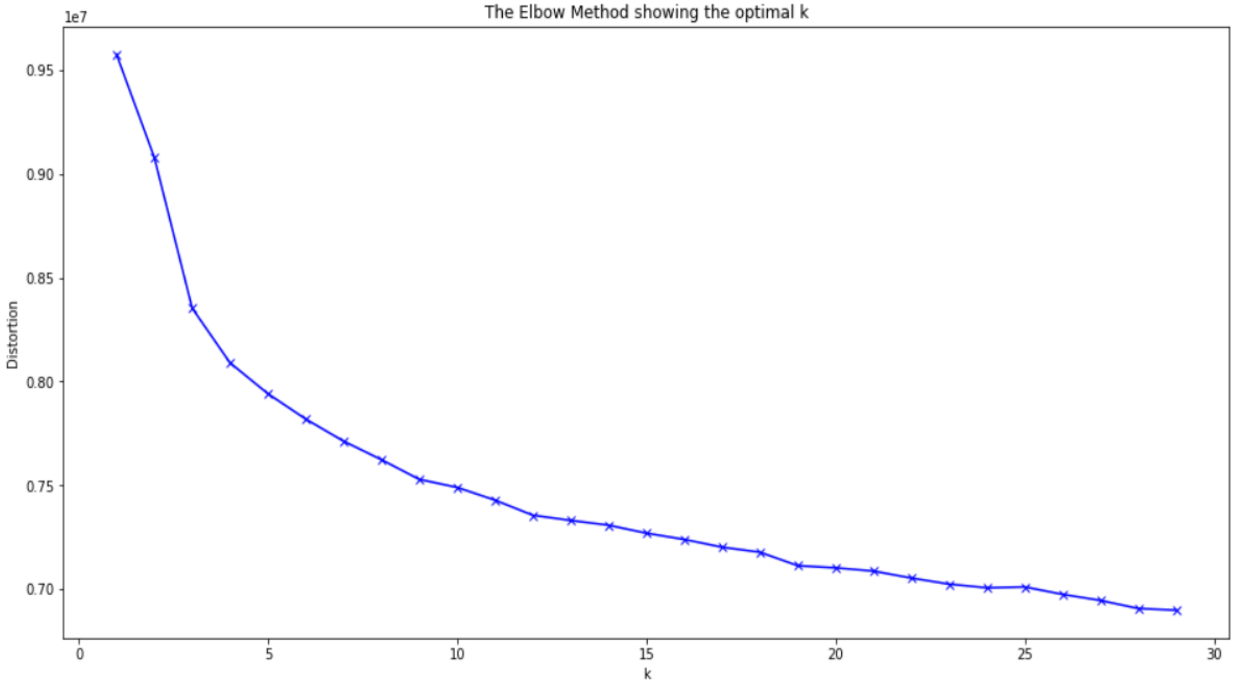
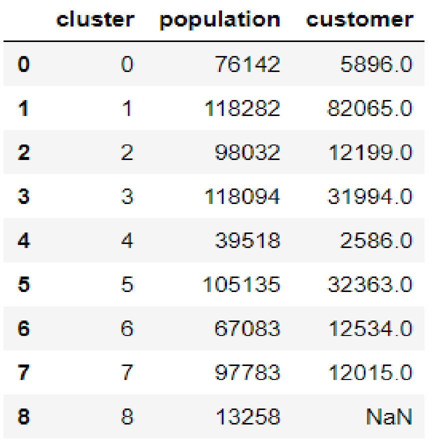


Figure 4: Elbow Method for KMeans Algorithm

From this plot I decided that the optimal number of clusters is 9 clusters, And then I printed the distribution of the general population and the customers in each cluster



	cluster	population	customer
0	0	76142	5896.0
1	1	118282	82065.0
2	2	98032	12199.0
3	3	118094	31994.0
4	4	39518	2586.0
5	5	105135	32363.0
6	6	67083	12534.0
7	7	97783	12015.0
8	8	13258	NaN

Figure 5: Population and customers Distribution among cluster

And then I plotted both the populations and got this result

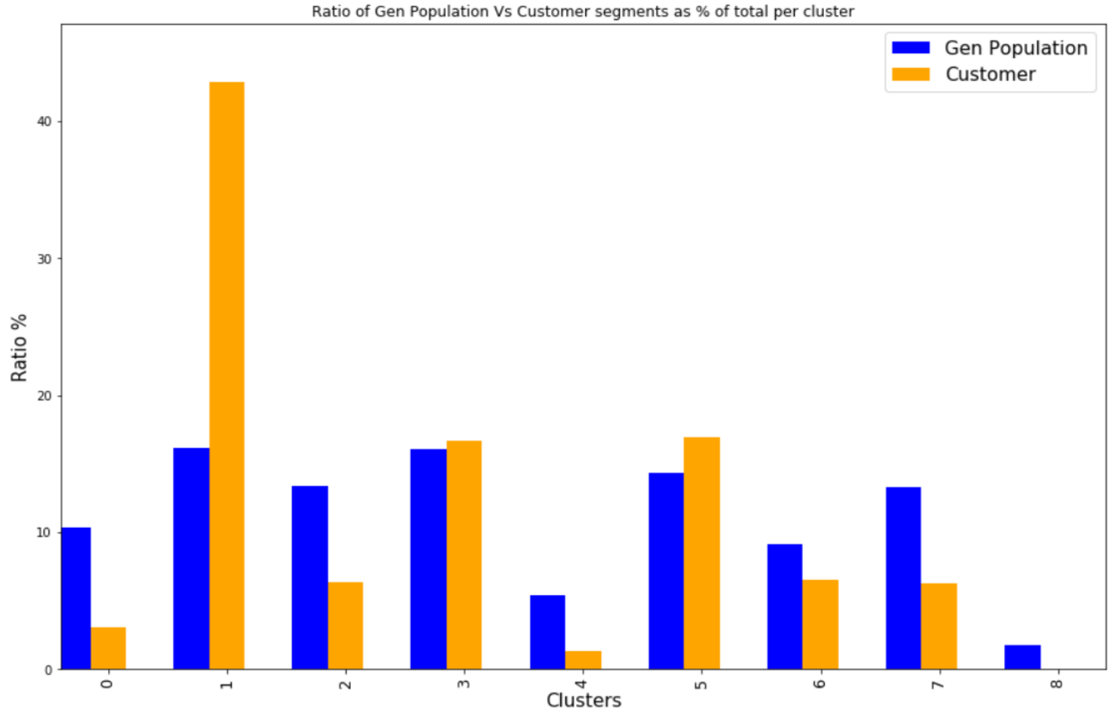
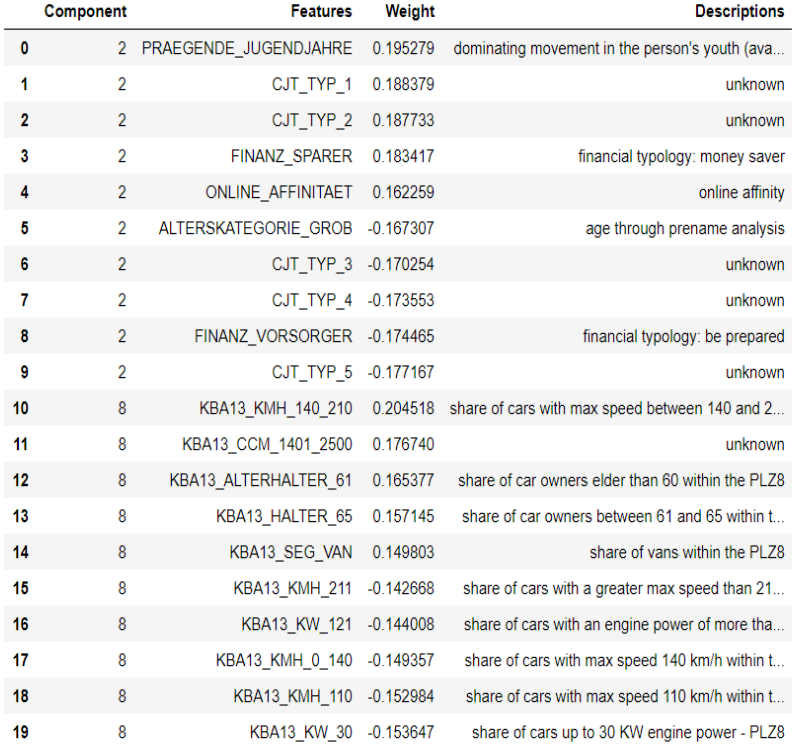


Figure 6: customer and general population Clusters

Key findings:

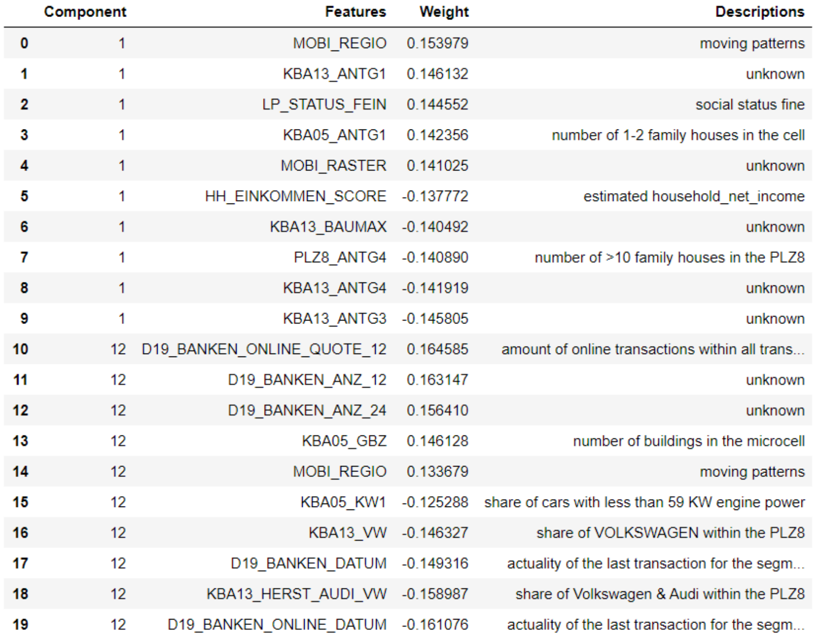
People in Clusters 1,3, and 5 are the targets for the company especially cluster 1, So now I have to discover what components are related to these clusters and what are the similarities between these customers and the general population.

Clusters analysis:



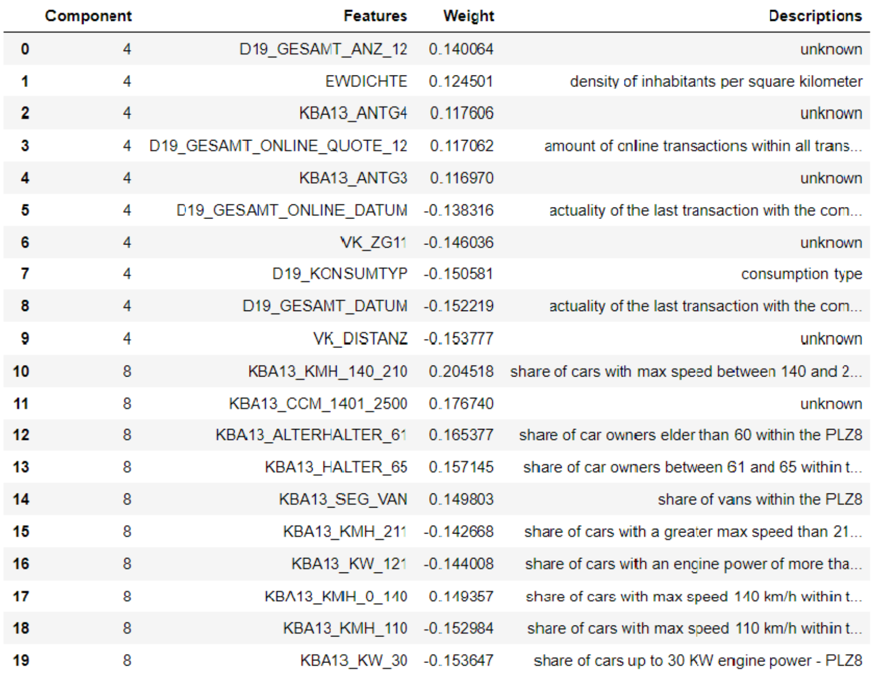
	Component	Features	Weight	Descriptions
0	2	PRAEGENDE_JUGENDJAHRE	0.195279	dominating movement in the person's youth (ava...
1	2	CJT_TYP_1	0.188379	unknown
2	2	CJT_TYP_2	0.187733	unknown
3	2	FINANZ_SPARER	0.183417	financial typology: money saver
4	2	ONLINE_AFFINITAET	0.162259	online affinity
5	2	ALTERSKATEGORIE_GROB	-0.167307	age through prename analysis
6	2	CJT_TYP_3	-0.170254	unknown
7	2	CJT_TYP_4	-0.173553	unknown
8	2	FINANZ_VORSORGER	-0.174465	financial typology: be prepared
9	2	CJT_TYP_5	-0.177167	unknown
10	8	KBA13_KMH_140_210	0.204518	share of cars with max speed between 140 and 2...
11	8	KBA13_CCM_1401_2500	0.176740	unknown
12	8	KBA13_ALTERHALTER_61	0.165377	share of car owners elder than 60 within the PLZ8
13	8	KBA13_HALTER_65	0.157145	share of car owners between 61 and 65 within t...
14	8	KBA13_SEG_VAN	0.149803	share of vans within the PLZ8
15	8	KBA13_KMH_211	-0.142668	share of cars with a greater max speed than 21...
16	8	KBA13_KW_121	-0.144008	share of cars with an engine power of more tha...
17	8	KBA13_KMH_0_140	-0.149357	share of cars with max speed 140 km/h within t...
18	8	KBA13_KMH_110	-0.152984	share of cars with max speed 110 km/h within t...
19	8	KBA13_KW_30	-0.153647	share of cars up to 30 KW engine power - PLZ8

Figure 7: Cluster 1 components, weights and Descriptions



	Component	Features	Weight	Descriptions
0	1	MOBI_REGIO	0.153979	moving patterns
1	1	KBA13_ANTG1	0.146132	unknown
2	1	LP_STATUS_FEIN	0.144552	social status fine
3	1	KBA05_ANTG1	0.142356	number of 1-2 family houses in the cell
4	1	MOBI_RASTER	0.141025	unknown
5	1	HH_EINKOMMEN_SCORE	-0.137772	estimated household_net_income
6	1	KBA13_BAUMAX	-0.140492	unknown
7	1	PLZ8_ANTG4	-0.140890	number of >10 family houses in the PLZ8
8	1	KBA13_ANTG4	-0.141919	unknown
9	1	KBA13_ANTG3	-0.145805	unknown
10	12	D19_BANKEN_ONLINE_QUOTE_12	0.164585	amount of online transactions within all trans...
11	12	D19_BANKEN_ANZ_12	0.163147	unknown
12	12	D19_BANKEN_ANZ_24	0.156410	unknown
13	12	KBA05_GBZ	0.146128	number of buildings in the microcell
14	12	MOBI_REGIO	0.133679	moving patterns
15	12	KBA05_KW1	-0.125288	share of cars with less than 59 KW engine power
16	12	KBA13_VW	-0.146327	share of VOLKSWAGEN within the PLZ8
17	12	D19_BANKEN_DATUM	-0.149316	actuality of the last transaction for the segm...
18	12	KBA13_HERST_AUDI_VW	-0.158987	share of Volkswagen & Audi within the PLZ8
19	12	D19_BANKEN_ONLINE_DATUM	-0.161076	actuality of the last transaction for the segm...

Figure 8: Cluster 3 components, weights and Descriptions



	Component	Features	Weight	Descriptions
0	4	D19_GESAMT_ANZ_12	0.140064	unknown
1	4	EWDICHTE	0.124501	density of inhabitants per square kilometer
2	4	KBA13_ANTG4	0.117606	unknown
3	4	D19_GESAMT_ONLINE_QUOTE_12	0.117062	amount of online transactions within all trans...
4	4	KBA13_ANTG3	0.116970	unknown
5	4	D19_GESAMT_ONLINE_DATUM	-0.138316	actuality of the last transaction with the com...
6	4	VK_ZG11	-0.146036	unknown
7	4	D19_KONSUMTYP	-0.150581	consumption type
8	4	D19_GESAMT_DATUM	-0.152219	actuality of the last transaction with the com...
9	4	VK_DISTANZ	-0.153777	unknown
10	8	KBA13_KMH_140_210	0.204518	share of cars with max speed between 140 and 2...
11	8	KBA13_CCM_1401_2500	0.176740	unknown
12	8	KBA13_ALTERHALTER_61	0.165377	share of car owners elder than 60 within the PLZ8
13	8	KBA13_HALTER_65	0.157145	share of car owners between 61 and 65 within t...
14	8	KBA13_SEG_VAN	0.149803	share of vans within the PLZ8
15	8	KBA13_KMH_211	-0.142668	share of cars with a greater max speed than 21...
16	8	KBA13_KW_121	-0.144008	share of cars with an engine power of more tha...
17	8	KBA13_KMH_0_140	-0.149357	share of cars with max speed 140 km/h within t...
18	8	KBA13_KMH_110	-0.152984	share of cars with max speed 110 km/h within t...
19	8	KBA13_KW_30	-0.153647	share of cars up to 30 KW engine power - PLZ8

Figure 9: Cluster 5 components, weights and Descriptions

After clusters analysis the company should target people with the following traits:

- Middle class people who tends to save money and may invest with it
- People who likely use moderated cars or vans with maximum speed between (140-210 Km/h)
- Car owners elder than 60 and their age ranges between (61-65)
- People in moderate neighborhoods with 1-2 families in the house
- Their social status is fine
- Their online transactions are within segment banking
- Their financial topology is less likely to be prepared
- People who less likely to own cars with high engine power
- People to less likely to do their segment banks transactions online

Supervised learning:

In this part of the project I have to build a supervised model to predict if the person is customer or not based on the demographic features using the mailout dataset which have the same features of the azdias dataset but with extra column named ‘RESPONSE’ which have label 1 if person is customer and 0 if not, so I have used the same preprocessing steps I used with azdias dataset on the mailout dataset.

Models and algorithms:

Benchmark model:

I tried many models to select the best model with the best performance, so I decided to select Random forest model as the benchmark model because it is not a slow model to train and it performs well with large datasets, I got a score of 0.51 which is bad but the other models performed better.

Other models\Algorithms used:

Model	Score
Logistic Regression	0.66
Adaboost	0.73
GradientBoosting	0.75
Naïve Bayes	0.61

Adaboost and gradientBoosting Classifiers got the best results and their scores were close to each other but the gradientBoosting classifier took about twice the training time of the adaboost classifier, So I decided to do hyperparameter tuning to both the models using GridSeardcv.

Hyperparameter Tuning:

Adaboost Classifier:

After reading the scikit learn article and some other articles from the internet I found that the most valuable parameters to be altered are the learning rate and n_estimators, after trying several values I found that the best values for it are :

- Learning rate: 0.1
- N_estimators: 50

And I got a score of 0.764 which is a good score.

GradientBoost Classifier:

I decided to try different sets of these parameters:

- Learning rate
- N_estimators
- Max_depth
- Min_samples_split

And found that the best values for them are:

- Learning rate:0.01
- N_estimator:50
- Max_depth:5
- Min_samples_split:4

And I got a score of 0.759 which is not a big difference from the basic model parameters.

Results

I started to do the same preprocessing steps I did earlier on the test dataset then used the best model for both the classifiers (Adaboost,GradientBoost) and I got a better result for the adaboost classifier which is 0.80337, while the gradient boost scored 0.79416.

By the Adaboost scoring 0.80337 I got the 50th position on the Kaggle competition. The second top score is 0.81063 which is not that far.

Future improvement steps:

- Understanding more features and do better and do deeper feature engineering
- Try to balance the training dataset classes by doing up-sampling or down-sampling
- Onehot encode categorical features

Important Notes:

- Before dropping the 'CAMEO_DEU_2015' I tried to one-hot encode it using pandas get_dummies function but it didn't provide any improvement in the accuracy of the model and instead increase the columns of the dataset, so I decided to remove it.
- I tried svm and xgboost models but I got bad results for the basic parameters for them and it took so much time training the model, then tried to hyper tune their parameters and also got bad results so I decided to remove them.
- When I used PCA with the mailout-train dataset before training, the performance degraded so I decided to not use it with the supervised learning models.
- When I removed the 30% of rows in the customers dataset that contains more than 16 missing values the performance degraded so I didn't remove any rows from this dataset.
- I tried using minmax scaler instead of standard scaler but it slightly degraded the performance

References

- https://chrisalbon.com/machine_learning/feature_selection/drop_highly_correlated_features/
- https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/