

# Machine Learning Capstone Project

*DATE: Feb 23, 2020*

*MACHINE LEARNING ENGINEER*

*NANODEGREE*

*JAHID AHSAN KHAN*

# Contents

Definition .....	3
Project Overview .....	3
Problem Statement .....	4
Metrics.....	4
Analysis .....	4
Data Exploration & Visualization.....	4
• Dataset Exploration.....	4
• Missing Value Analysis .....	5
• Data Types Analysis.....	6
Algorithms and Techniques.....	6
• Unsupervised Learning Model .....	6
• Supervised Learning Model .....	7
Benchmark .....	7
Methodology .....	8
Data Preprocessing .....	8
Implementation.....	9
• Unsupervised Learning Model .....	9
• Supervised Learning Model .....	14
Refinement.....	14
Results.....	14
Model Evaluation and Validation.....	14
Justification .....	15
Conclusion .....	16
Reflection & Improvements .....	16

# Definition

## Project Overview

Acquiring customers more efficiently is an essential part of running a successful company. One of the ways to improve its efficiency is through targeted marketing. For a long time, the process of choosing the most likely customers has largely been done using human intuitions and experiences. In the recent times, with the advancements in data collection techniques, it is now possible to use large amount of data to make more efficient and effective business decisions instead of solely relying on gut feelings. This project aims to utilize the data obtained from its customers and the population of Germany to find segments in population which are most likely to become customers for the company. The dataset is provided by Arvato Financial Solutions. The data provided is not publicly available and is allowed to be used only for this project. The provided data consists of the following files.

- 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).
- Udacity\_MAILOUT\_052018\_TRAIN.csv: Demographics data for individuals who were targets of a marketing campaign; 42 982 persons (rows) x 367 (columns).
- Udacity\_MAILOUT\_052018\_TEST.csv: Demographics data for individuals who were targets of a marketing campaign; 42 833 persons (rows) x 366 (columns).

There are two additional files along with the data files as follows.

- DIAS Information Levels - Attributes 2017: Contains Information level, attributes and description of columns in dataset.
- DIAS Attributes - Values 2017: Contains attributes, description, value and meaning of these values.

Each row of the demographics files represents a single person, but also includes information outside of individuals, including information about their household, building, and neighborhood.

The "CUSTOMERS" file contains three extra columns ('CUSTOMER\_GROUP', 'ONLINE\_PURCHASE', and 'PRODUCT\_GROUP'), which provide broad information about the customers depicted in the file. The original "MAILOUT" file included one additional column, "RESPONSE", which indicated whether or not each recipient became a customer of the company. For the "TRAIN" subset, this column has been retained, but in the "TEST" subset it has been removed.

## Problem Statement

From a business perspective, the underlying question is: “How can a mail order company acquire new clients more efficiently?”

The Arvato Financial Solution has tasked us to identify and analyze the traits that its customers share with the population in such a way that one can identify which segments of population will be most likely to become customers in future. The goal is to identify such segments of the population which can then be used for targeted marketing campaigns to achieve a higher expected rate of return.

The project is divided into three main parts. In the first part, the unsupervised learning model is required to analyze the attributes of established customers and the general population in order to create customer segments. For the second part, we are required to train a supervised learning model based on attributes of targets of mail order campaign provided in third dataset which we will then use to predict the responses of individuals in future campaigns. In the last part, we will tune our model to predict the responses of individuals with given attributes in test dataset and will submit it in Kaggle competitions page where our model will compete against models of other users. The one with the highest score based on the AUROC (Area Under Receiver Operative Characteristic) curve metric will have the best predicting model.

## Metrics

Since the data is highly imbalanced, the accuracy score will not determine good results. Therefore, the Area under the ROC curve (AUC) will be used as a metric as it ranks a randomly chosen positive instance higher than a randomly chosen negative example and is also used in Kaggle competitions. The score of AUROC curve ranges from 0 to 1 with 1 perfectly identifying every label correctly and 0 perfectly misidentifying every label. The score of 0.5 shows that a model cannot distinguish between negatives and positives at all.

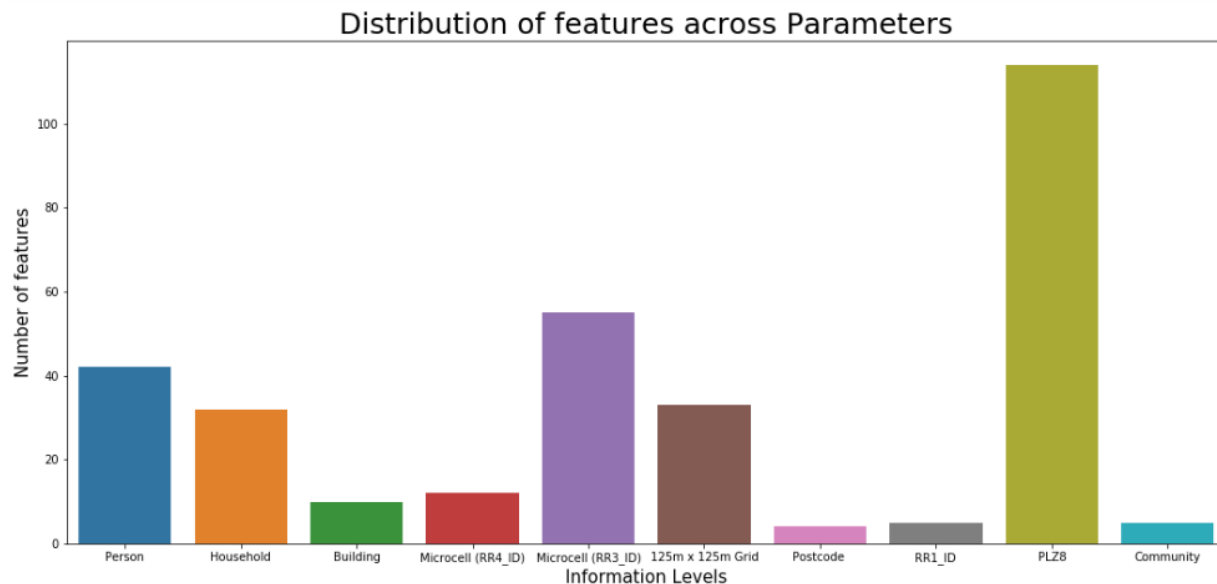
## Analysis

### Data Exploration & Visualization

- *Dataset Exploration*

The dataset of population of Germany contains 891,211 rows and 366 features while the dataset of customers contains 191,652 rows and 369 features. The names of the features are not self-explanatory hence, the given additional files are referred to understand the terminologies used. The features cover wide variety of parameters including features related to individual (e.g. age, gender, financial type), household (e.g. estimated household net income,

main age within household, transaction activity), and vehicle related information (e.g. share of car owners less than 31y.o, share of cars less than 1399 cc, share of upper class cars). The graph below shows the number of features covered by different parameters.

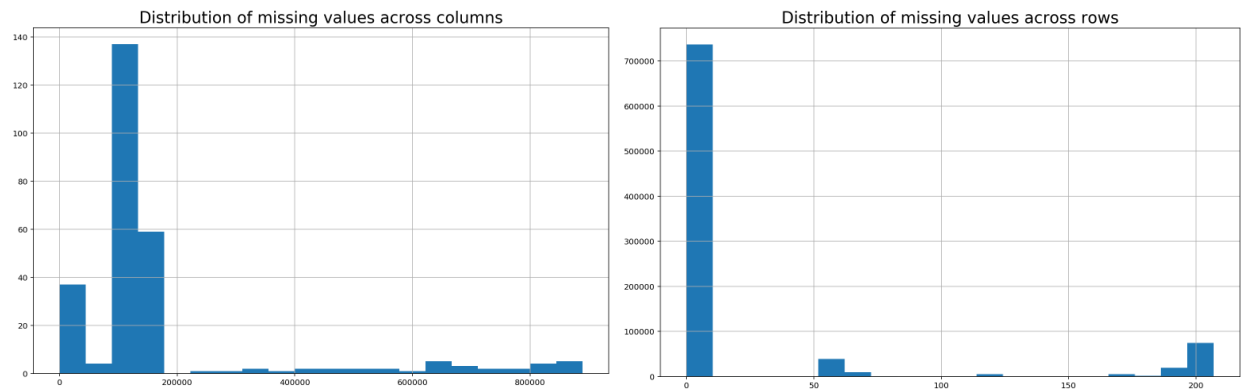


- *Missing Value Analysis*

The data contains many missing values as well as values indicating missing information. A new dataset was build based on missing information values (-1, 0, 9) mentioned in DIAS Attributes dataset. This new dataset was then used to re-encode the matched values in population dataset to NaN values. The head of the new dataset with first 10 rows is shown below.

	Attributes	missing_or_unknown
0	AGER_TYP	[-1, 0]
1	ALTERSKATEGORIE_GROB	[-1, 0]
2	ALTER_HH	[0]
3	ANREDE_KZ	[-1, 0]
4	BALLRAUM	[-1]
5	BIP_FLAG	[-1, 0]
6	CAMEO_DEUG_2015	[-1]
7	CAMEO_DEUINTL_2015	[-1]
8	CJT_GESAMTTYP	[0]
9	D19_BANKEN_ANZ_12	[0]

The population dataset was then plotted based on missing values in columns as well as in rows.



- *Data Types Analysis*

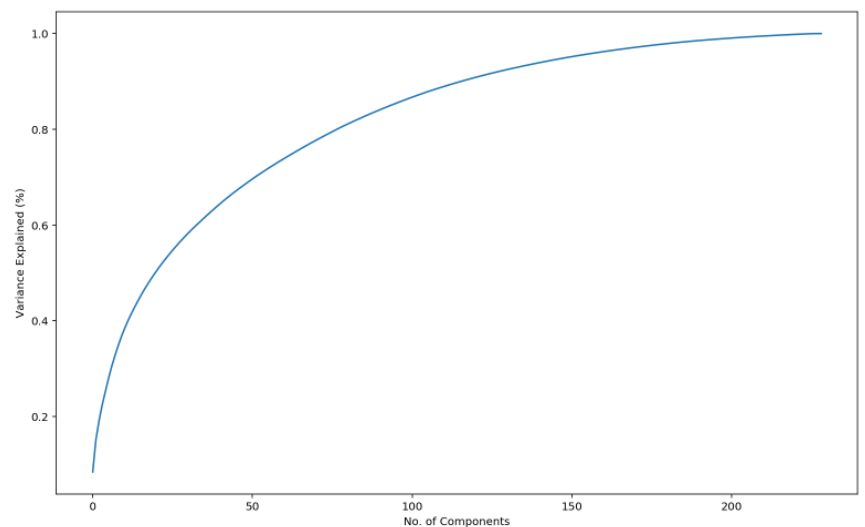
While the majority of data is represented in numerical form, the dataset contains multiple data types where categorical data was encoded later using dummy variables and mixed type variables and categorical variables with greater than 10 different values were dropped for simplicity.

## Algorithms and Techniques

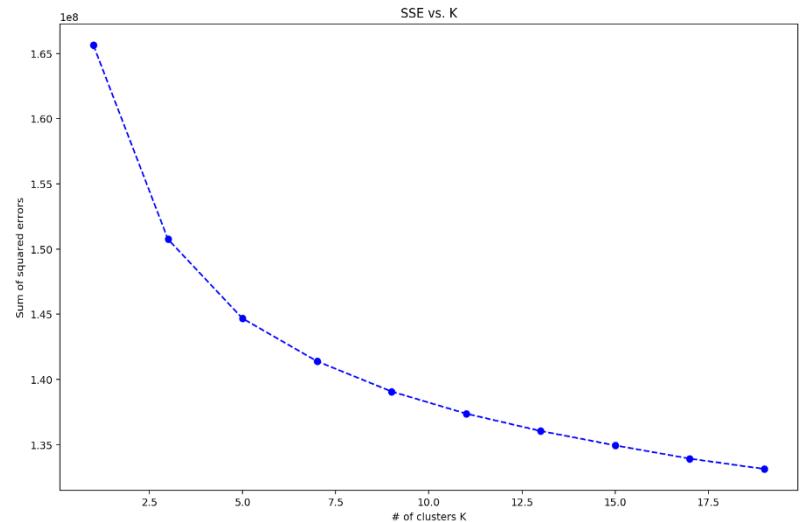
- *Unsupervised Learning Model*

The main bulk of the analysis lies in this part of the project. Here, we used unsupervised learning techniques to describe the relationship between the demographics of the company's existing customers and the general population of Germany. By the end of this part, we were able to describe parts of the general population that are more likely to be part of the mail-order company's primary customer base, and which parts of the general population are less so.

Before applying unsupervised learning algorithm, the dimensionality reduction techniques were applied with Principle Component Analysis (PCA) to reduce the number of features within dataset while still retaining 90% of variance. The number of components were chosen by looking at the graph. Based on this graph, it was decided to choose 125 components which explained more than 90% variance in data.



After applying dimensionality reduction techniques, K-means clustering method was used to identify segments of population which are most likely to become customers. In order to identify which number of clusters will be optimal, we created the elbow plot by plotting K-means cluster models from 1 to 20 against sum of squared error. Based on this graph, it was opted to go for 7 clusters as the curve started to flatten after that. We then used this model for further analysis on population and customers data.



- *Supervised Learning Model*

To predict the likelihood of an individual to reply to the mailing campaign, various boosting algorithms were deployed and their best scores were compared against each other to choose the best scoring model for further tuning. The models deployed included AdaBoostClassifier, GradientBoostingClassifier, XGBoost Classifier, CatBoost Classifier, and LGBM Classifier. Based on the scores, XGBoost model was selected and further tuned using RandomizedSearchCV to give the best predicting results.

Model	Best Score
AdaBoostClassifier	0.7259
GradientBoostingClassifier	0.7531
XGBoost Classifier	0.7574
CatBoost Classifier	0.7335
LGBM Classifier	0.7087

## Benchmark

The model chosen for benchmark was XGBoost base model. Since XGBoost model handles missing values on its own, it was deployed on training data without preprocessing and then the predictions for test data were submitted on competitions page to get the benchmark score. The benchmark model score to beat under the metric AUROC was 0.79772.

Name	Submitted	Wait time	Execution time	Score
submission_kaggle.csv	just now	0 seconds	0 seconds	0.79772

Complete

[Jump to your position on the leaderboard](#) ▼

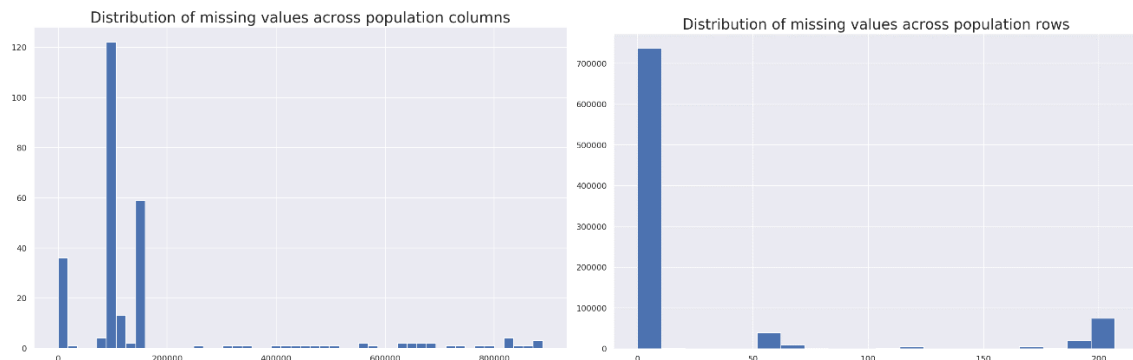
# Methodology

## Data Preprocessing

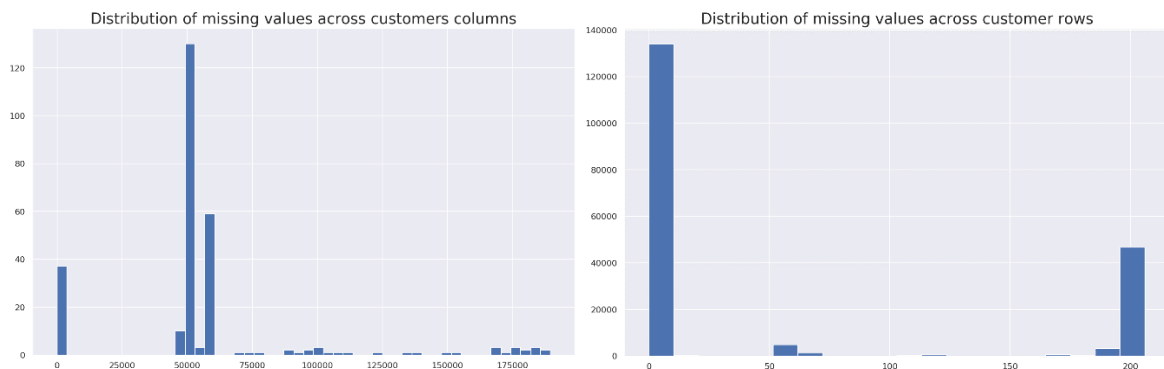
The data preprocessing involved the following steps:

- 1) The “DIAS Attributes – Values” dataset was used to remap values in population and customers data that represent missing data as NaNs.
- 2) Since the feature labels and their values were not self-explanatory, only the features whose description were available in the abovementioned dataset were retained while the rest were dropped in the unsupervised learning part.
- 3) The integer values were converted to float values in population and customers dataset.
- 4) Based on distribution of data using exploratory data analysis, the columns with greater than 200,000 missing values and rows with greater than 50 missing values were dropped in population dataset.
- 5) Similar to above, in customers dataset, the columns with greater than 65,000 missing values and rows with greater than 50 missing values were dropped.

### Population data



### Customers Data



The below mentioned steps were applied simultaneously on population and customers dataset.

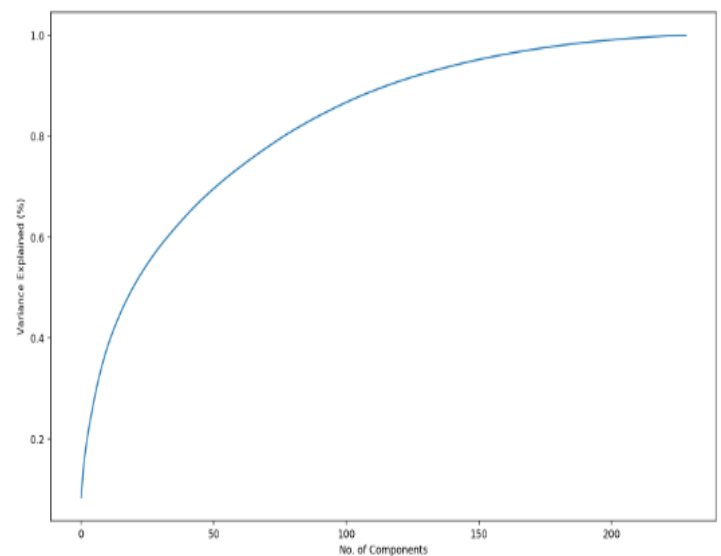


- 6) The only binary variable found “OST\_WEST\_KZ” was re-encoded as 0 and 1.
  - 7) The features with greater than 10 levels were dropped for simplicity.
  - 8) The remaining categorical features were re-encoded using one hot encoding and the unnecessary columns were dropped.
  - 9) The missing values were then imputed using mean strategy.
  - 10) The rows containing outliers were removed using z-score.
  - 11) The data was then normalized to zero mean and unit variance using “StandardScaler”.
  - 12) In the supervised learning part, no columns were dropped based on missing values and only the columns with mixed value types were dropped for the model to run.
- Furthermore, the data was not normalized before training as the selected model (XGBoost) scored better using non-normalized data. The rest of pre-processing on “train” and “test” dataset were same as above.

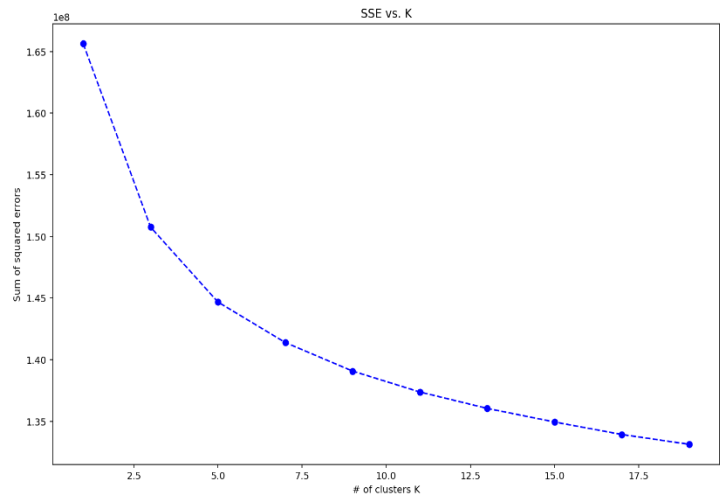
## Implementation

- *Unsupervised Learning Model*

After the data preprocessing steps were implemented, it was now ready for training. But due to large number of features, the process of training would take a lot longer. Therefore, we first applied dimensionality reduction techniques to speed up the training process. for this task, we used Principle Component Analysis (PCA) method from Scikit-Learn. To choose the total number of components, we started by fitting PCA on population dataset and then plotted the explained variance ratio of all principle components. Using this plot, we decided to keep the first 120 principle components that explained more than 90% cumulative variance in the data.



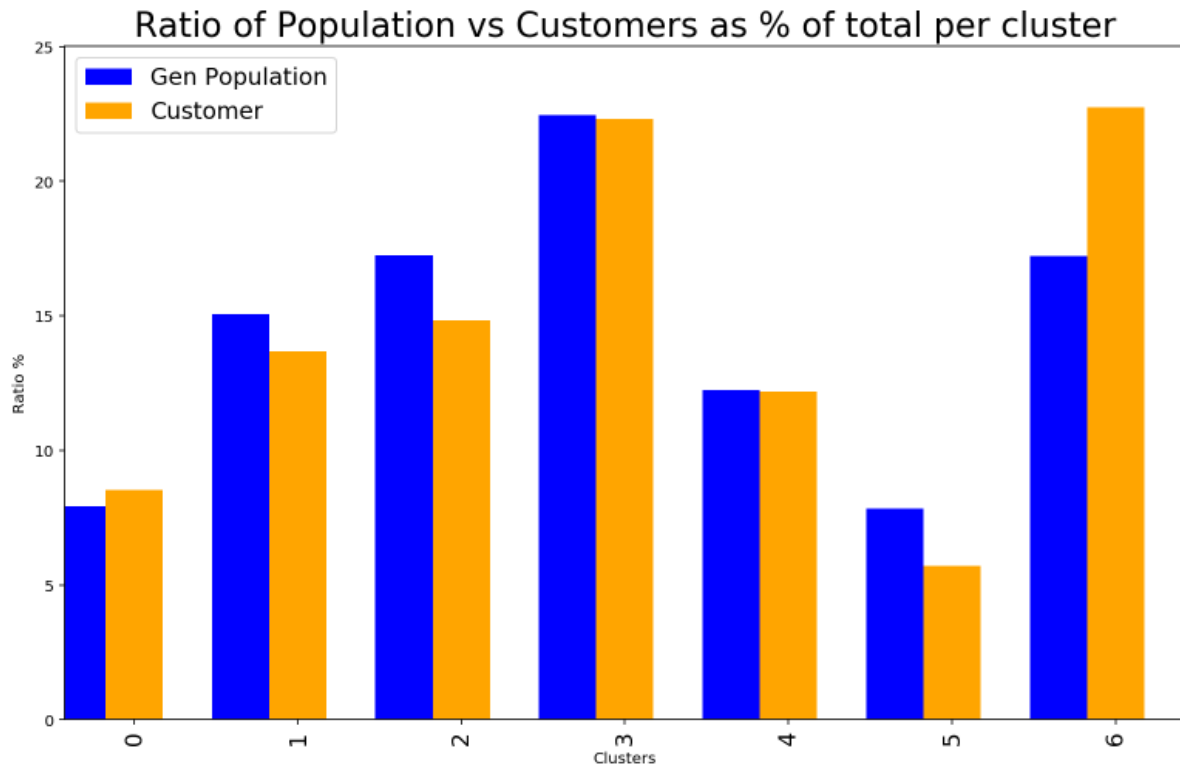
After the total number of features were reduced to 120, we now applied K-Means algorithm to segment the data into clusters. To find the optimal number of clusters, the elbow plot was built by plotting number of clusters against the sum of squared errors. According to the elbow method – which is one of the most popular methods of determining the optimal number of clusters “k” – we have to select the value of k at the “elbow” i.e. the point after which the error starts decreasing in a linear fashion. Based on our graph, it was decided to choose the value of k=7.



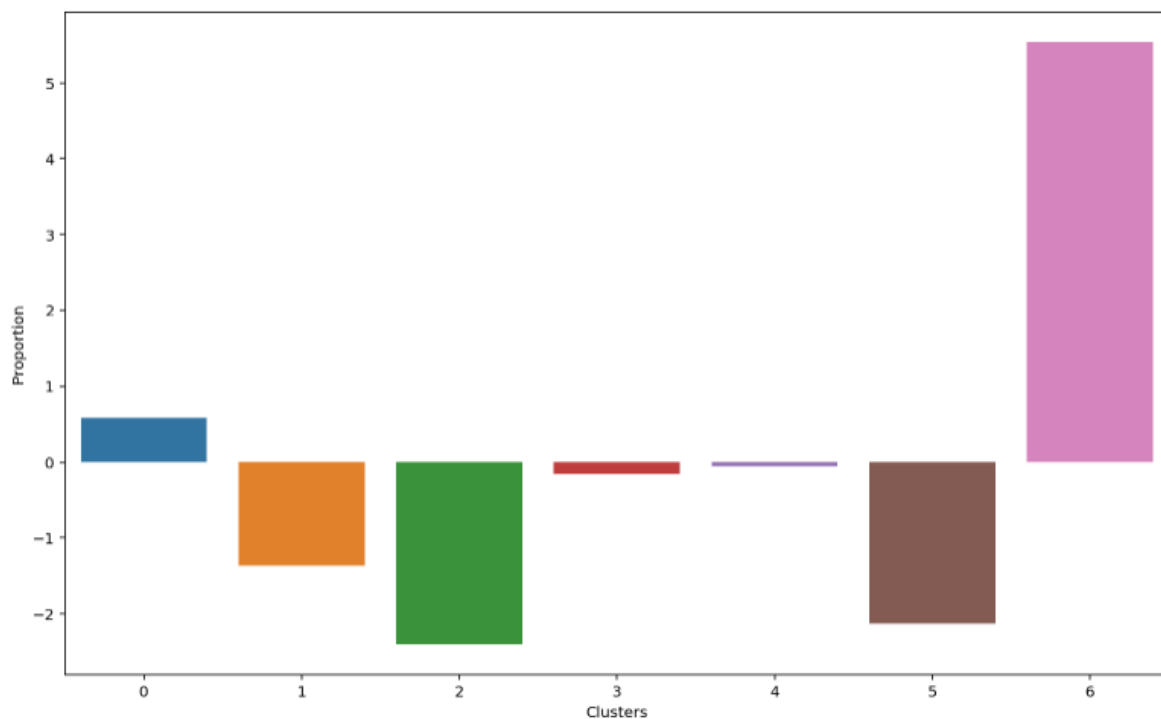
After the “k” was selected, we fit the K-Means model on population dataset and then used the model to obtain cluster predictions on both population and customer dataset. We then used these predictions to create a new dataset representing the value counts of both the population and customers in each cluster along with their difference in ratio of percentages of values in each cluster.

Cluster	Customer	Population	Difference
0	11007	57698	0.575814
1	17691	109405	-1.369292
2	19172	125248	-2.404062
3	28823	163161	-0.154085
4	15731	88825	-0.053135
5	7389	57006	-2.128953
6	29400	125087	5.533713

Furthermore, the ratio of population and customers as percentage of total values per cluster was plotted. Along with it, the difference column was also plotted to search for over-represented and under-represented clusters.



Based on the graph below, we can see what kind of people are part of a cluster that is over-represented in the customers data (compared to population data) i.e. cluster 6 and what kind of people are part of a cluster that is most under-represented in the customer data (compared to the population data) i.e. cluster 2.

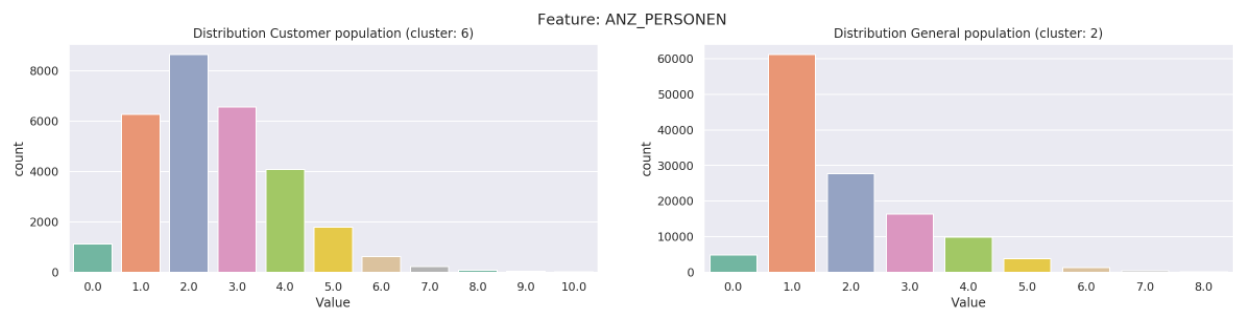


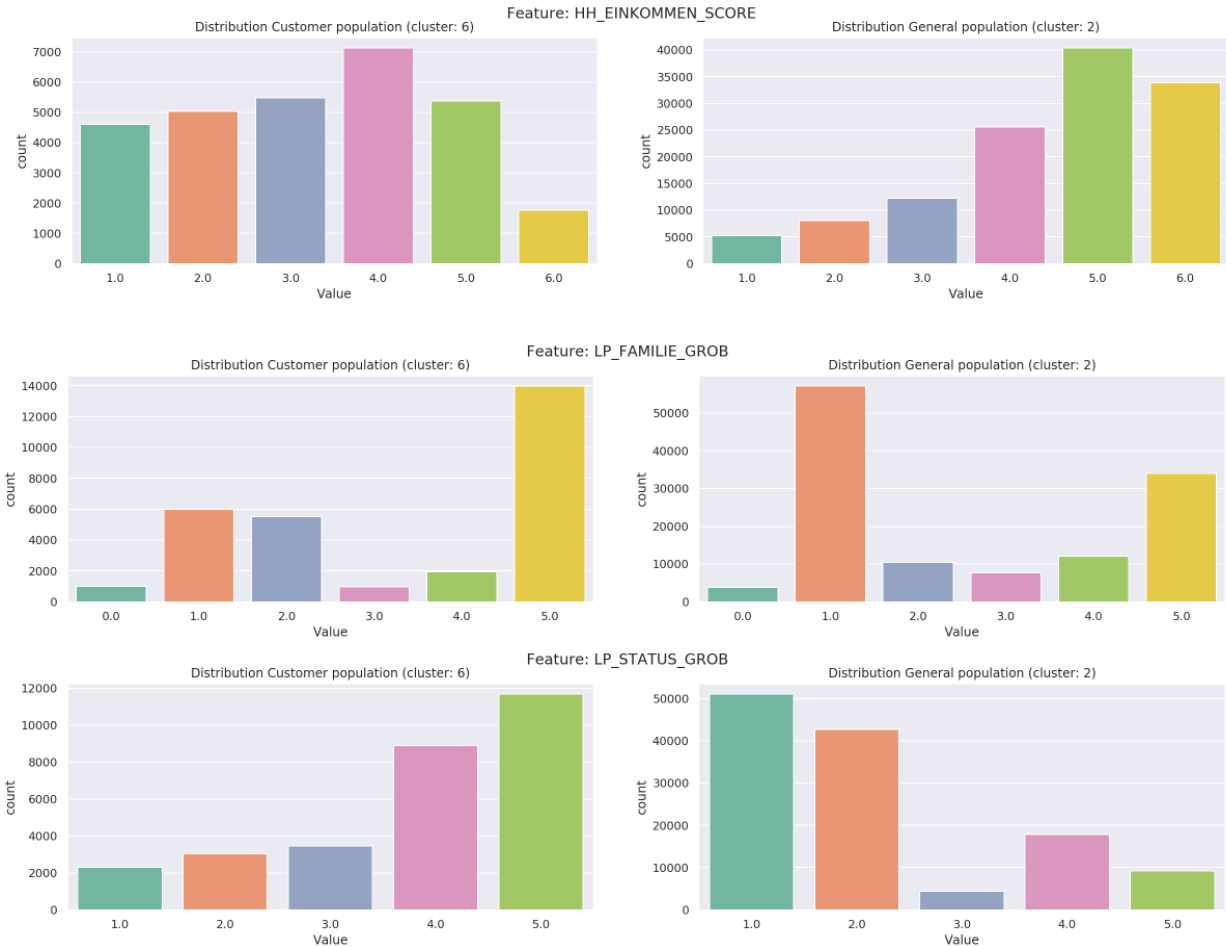
By comparing the features in over-represented and the most under-represented clusters, following observations were made.

The over-represented cluster was composed largely of males aged 45 and above (ANREDE\_KZ feature representing gender & ALTERSKATEGORIE\_GROB feature representing age classification) while the under-represented cluster had slight majority of females aged mostly below 45 years.

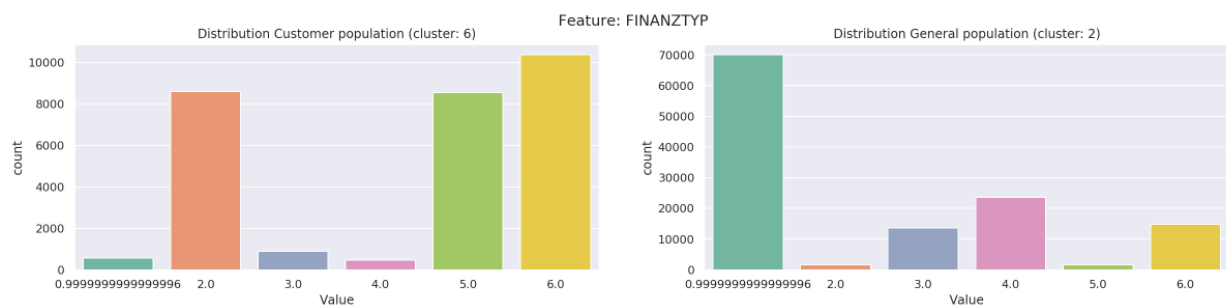


In terms of household, the average number of adult persons in household in over-represented cluster were 2 with a net income ranging from high income to low income, the family system was single parent with social status average earners. On the other hand, in under-represented cluster, the average number of adult persons in household was 1 with net income ranging from low income to very low income, the family system was single with social status low income earners. (ANZ\_PERSONEN feature representing adult persons in household, HH\_EINKOMMEN\_SCORE feature representing est. household net income, LP\_FAMILIE\_GROB feature representing family type, & LP\_STATUS\_GROB feature representing social status).





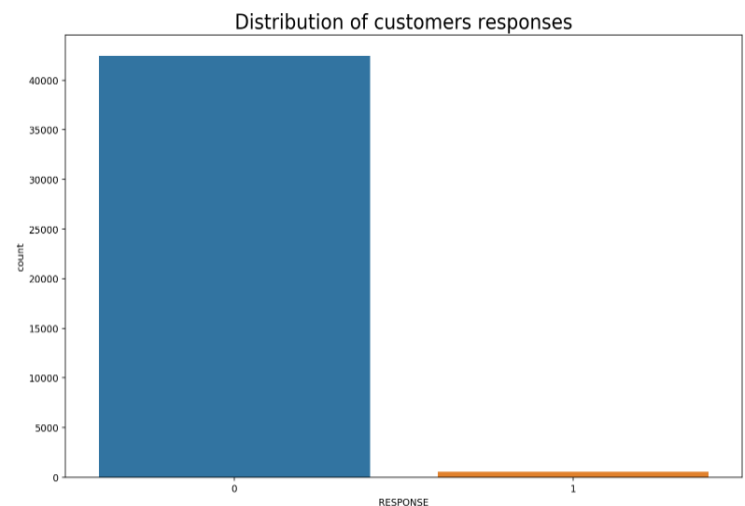
In addition to that, the best describing feature for the financial type was “FINANZTYP” with persons in over-represented cluster having “unremarkable”, “investor”, and “money saver” financial type while the under-represented cluster having “low financial interest”.



Lastly, when it comes to individual traits and behaviors, the individuals in over-represented cluster have on average higher affinity of having a fight full attitude, being critical minded, being material minded, being traditional minded, and being religious minded than individuals in under-represented cluster (SEMIO\_KAEM, SEMIO\_KRIT, SEMIO\_MAT, SEMIO\_PFLICHT, SEMIO\_REL features representing fight full attitude, critical mindedness, material mindedness, traditional mindedness, religiosity).

- *Supervised Learning Model*

In this part, the training and testing dataset was preprocessed using the techniques mentioned in data preprocessing part. In addition to the similar columns as in population and customers dataset, the training dataset contained one more column i.e. RESPONSE column indicating which individuals became customers of the company. The distribution of responses in the dataset can be seen in the graph shown. Here one can see that the responses are very imbalanced. Hence, the use of accuracy as a measure would have been useless and instead the AUROC metric was used for model training.



After the data preprocessing function was performed, the dataset was fit through multiple boosting algorithms using GridSearchCV and default parameters. The algorithms used included AdaBoostClassifier, GradientBoostingClassifier, XGBoost Classifier, CatBoost Classifier, and LGBM Classifier. As mentioned above, the XGBoost model outscored other models and was selected for hyperparameter optimization process.

## Refinement

One of the important steps in achieving a better score with machine learning model lies in data preprocessing and feature engineering techniques. Upon iteration, I found that by not dropping columns with excess NaN values like I did during unsupervised learning part, it resulted in better score. Furthermore, I also found that the “mean” imputation strategy gave a better score than either “median” or “most frequent”. Lastly, normalizing the model with StandardScaler using the same hyperparameters gave worse score than using the non-normalized model. Hence, the final model for tuning used the non-normalized data.

## Results

### Model Evaluation and Validation

The model was tuned using GridSearchCV with StratifiedKFold first with 5 splits and later with 10 splits as sampling method. The total search hyperspace used for optimization of the model is as follows.

"n\_estimators": [20,50,100,1000]

"max\_depth": [2,3,4,5,6,7]

"learning\_rate": [0.1,0.05,0.001]

"subsample": [0.8,1]

"scale\_pos\_weight": [10,25,100]

"gamma": [0,1,5,10]

"colsample\_bytree": [0.3,0.5,0.8]

After first few iterations, I went with default values of some hyperparameters including "subsample" and "gamma" to reduce the training time. The best score I got on the validation set was 0.770858 using 10 splits as sampling method but the best score I got on leaderboard i.e. 0.80453 was obtained when the score on validation set was 0.766173 using sampling method of 5 splits. My overall best score on leaderboard was obtained with the following hyperparameters.


"colsample\_bytree": 0.5,

"learning\_rate": 0.01,

"max\_depth": 5,

"n\_estimators": 100,

"scale\_pos\_weight": 10

22	Jahid Ahsan		0.80453	23	2h
----	-------------	---	---------	----	----

## Justification

After I got the score of 0.80453 on leaderboard. I decided to continue training with different parameters and techniques which did give me a better score on validation set but failed to improve my score further on leaderboard. Moreover, since I already beat my benchmark score, therefore, I decided to stop here as the cross-validation process was very time consuming as well.

## Conclusion

### Reflection & Improvements

This was my first big project of its kind and really made me learn a lot about machine learning and its applications. One thing I did struggle during my work was with limited hardware access which forced me to shift my project to cloud. As a result, I ended up using Google Colab as it provided a Jupyter-esque environment with free CPU, RAM, & GPU access. But the downside of using Colab was strict idle timeouts with unexpected runtime disconnects and maximum run time of 12 hours which in fact at one point made it impossible for me to use GridSearchCV and eventually made me to switch towards RandomizedSearchCV. Due to these reasons, I also ended up only choosing XGBoost for hyperparameter tuning even though the base scores of GradientBoostingClassifier from Scikit-learn and CatBoost Classifier were also impressive. Overall, I have learned a great deal from this project and now aim to apply what I have learned here onto future projects.