Customer Segmentation Report for Arvato Financial Solutions

Data Scientist Nanodegree

Bassem Essam
October 2021

Definition

Project Overview

It's crucial for any company running any business related to customers to classify the customers and make segmentation for the groups of the customers. By understanding the behavior and the consumption pattern of the customers, the company can target the proper services and offers to the right customers.

One of the many applications of Machine Learning and Data Science is that the common patterns that can be detected between the data and computers can predict the response of the customers based on many variables that affect the behavior of the customers. Therefore, the company can make more profit and target the right customers without wasting the money and efforts to false audiences.

In this project, Arvato Financial Solutions provided a large database for about 900 thousand citizens in Germany. The database contains much information about the customers' social classes, consumption rates, how are their neighborhoods and their education, and so on. Using this data, we can find the common patterns between the customers and segment them. Then, this segmentation technique can be applied to a smaller data set of Arvato customers and find the special characteristics of the customers' database. This conclusion will be very helpful in understanding the customers and how to deal with them in any marketing campaign.

The next part of the project is to predict the response of targeted customers to a mail marketing campaign. In this part, we will create an ML model to efficiently predict how the customers respond to the mails.

In summary, the used datasets in this project are as follows.

- **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).

Problem Statement

The main business questions that need to be answered in this project are,

- What are the segments of the German population data set?
- How can these segments be applied to the Arvato customers database?
- What is the predicted response of the targeted customers to the mail marketing campaign?

To answer these questions, The Cross-Industry Standard Process for Data Mining (CRISP-DM) process will be followed in two parts with the following steps.

1. Business understanding.
2. Data understanding.
3. Data preparation.
4. Modeling.
5. Evaluation.
6. Deployment.

Part I: General Population and customers segmentation (Unsupervised Modeling)

In this part, The Demographics data for the general population of Germany (AZDIAS data set) and Demographics data for customers of a mail-order company (customers data set) will be used to make the segmentation of the data points. And the strategy that is planned to be used in this part is as follows.

1. **Data Understanding:** The features are in groups; each group of features represents an aspect of the population and the customers. The references and value attributes of each feature are provided in separate excel sheets.
2. **Data Preparation:** In this part, the common techniques of cleaning and preparing the data will be used as follows.
    - Handling missing values
    - Handling categorical features
    - Handling outliers
    - Data structure conversion
    - Feature engineering
3. **Modeling:** In this part, a data pipeline will be used with the following stages to build our unsupervised model.
    - Standard scaler transformer to scale the data.
    - Dimensions reduction, the PCA technique will be used to reduce the dimensions of the data. Also, the proper number of PCA components will be determined.
    - Unsupervised ML model, Kmeans algorithm will be used to segment the customers.

1. **Evaluation:** The evaluation of the KMeans model will be done using the elbow curve to determine the right number of clusters and fit the model with data to predict the labels of the data points. The same model will be used to predict the segments of the customers' data set.
2. **Deployment:** The deployment of the model can be done on a cloud-based platform or deploy with a web application, but in this project, we will focus on answering the business question and the deployment part will be done later.

Part II: Customers response prediction (Supervised Modeling)

In this part, we have already done most of the data cleaning/preparation work in Part I, therefore we will use the same assumptions and decisions made in the unsupervised modeling part. Then, the following steps will be taken to achieve the purpose of this part.

1. **Modeling:** In this stage, the techniques of imbalanced data sets should be followed as the response variable is imbalanced. Many model classifiers will be tested to get the model that has the best performance.

2. **Evaluation:** After choosing the model with the best performance, the evaluation will be done by predicting the response for the test dataset (which is provided without labels) and the results will be applied to the Kaggle competition.
3. **Deployment:** The deployment of this part will be integrated with the model in Part I.

Metrics

The metrics that will be used for the project will be divided into two parts as per the parts of the project.

1- **Unsupervised modeling:** The variance between the centroids and data points of each cluster will be calculated for each number of clusters used in the KMeans model and the elbow curve will illustrate the breakpoints which represent the best number of clusters.

2- **Supervised modeling:** The metrics in this part will be in two stages.

- Model selection: The model with the best performance to be selected will be determined by AUC (Area Under Curve) metric. Which represents the goodness of the model to distinguish between different classes. To achieve that ROC will be plotted for different models and the model with higher AUC will be used in the implementation.
- Model Deployment: In this section, the hyperparameters tuning will be done for the selected model and the metric to be used is AUC too. The reason the ROC curve is being used is that the data for the mail campaign is imbalanced and this makes the accuracy a bad choice for judging the model performance. On the other hand, precision and sensitivity can lead to the right decision.

Analysis

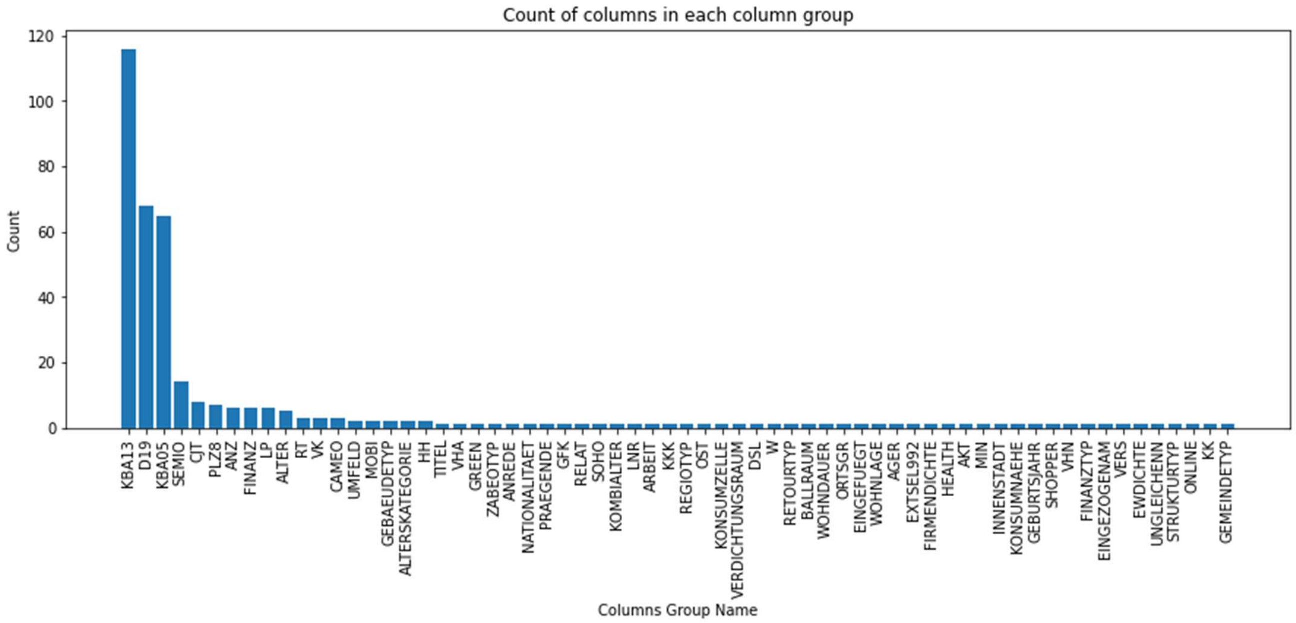
Data Exploration

In the data exploratory analysis, we see the columns are grouped in about 62 groups, each group describes a specific aspect. A summary of some of the highest groups and an explanation of some expressions used in the data set. is as follows.

- **PLZ8:** An area of about 500 households in the same area.
- **KBA13:** The information about the cars' owners and manufacturers in PLZ8 the customer belongs to.
- **D19:** Information about the bank transactions and the categories of payments of the customers.
- **KBA05:** The information about the cars' owners and manufacturers in the microcell the customer belongs to.
- **SEMIO:** Information about the social attributes of a person based on a customers survey.
- **CJT:** Customer-Journey-Typology relating to the preferred information and buying channels of consumers.
- **FINANZ:** Financial information about the customers.
- **LP:** (LEBENSPHASE) the life stage characteristics of the customers.
- **ALTER:** The age groups and children age groups of the customers.

Handling the columns in their group will make the data exploration and data preparation easier and interpretation of the data.

In the graph shown below, we can see the count of the groups in the data sets.



The following notes were noticed from the data exploration.

- Most of the features can be considered as categorical features as the values are nominal and represent a specific value.
- The following features have numeric values and these numbers represent their values. ANZ_HAUSHALTE_AKTIV, ANZ_HH_TITEL, ANZ_PERSONEN, ANZ_TITEL, GEBURTSJAHR, KBA13_ANZAHL_PKW und MIN_GEBAEUDEJAHR.
- EINGEZOGENAM is a date column that represents the date when the user added to data sets.
- CUSTOMER_GROUP, ONLINE_PURCHASE, PRODUCT_GROUP are three extra columns in customers' data sets that are not in the AZDIAS data set.

Methodology

Data Preprocessing

Part I Unsupervised Modeling:

In this section, we will investigate the flaws in the data and try to clean the data to be ready for modeling. The work on data processing will be done in the following sequence.

1- **Handling missing values:** missing values are the most effort aspect in cleaning these data sets as almost half of the columns with missing values. We have put into consideration if data are missing at random or not missing at random. Here is the summary of actions to be taken.
   - Some financial information can be not missing at random because customers might not be comfortable with telling about their sensitive information. In these cases, we can understand the relationships between different features, and the missing values can be imputed with the value corresponding to unknown if available.
   - In some cases, we impute the missing values with the mode value in the feature.

- If the percentage of the missing values is high, we can drop the column.

2- **Handling non-numeric features**

As mentioned earlier, most of the features are categorical, but some of the features have a string data type and hence, they should be handled in different methodology as follows.

- Firstly, we must impute the missing values as per the techniques mentioned above.
- Secondly, we should map the string values to a numeric value.
- If the percentage of missing values is high, the column will be dropped.
- It's important to note that mapping nominal values are not always a good idea as the numeric values can mislead the model predictions. In our case study, we can replace the categorical values with numbers as we are implementing an unsupervised model which means that the numeric data will not affect the data points and distances.

3- **Handling Outliers:**

There are no outliers detected in the datasets because most of the columns are categorical features with no values out of attributes provided in the supporting sheets.

4- **Handling Date Features:**

EINGEZOGENAM is a date column that represents the date when the user was added to the database, A reference date was taken to calculate the duration since the customer was added to the database.

Part II Supervised Modeling:

The same techniques and assumptions done in the previous section can be applied to the data sets of the mail campaign as the features are the same and the following differences are noticed between both data sets.
After imputing the missing values in mailout data sets, some of the columns that have not missed values in AZDIAS data sets. So, in the case the missing values are not high, we can drop these data points.

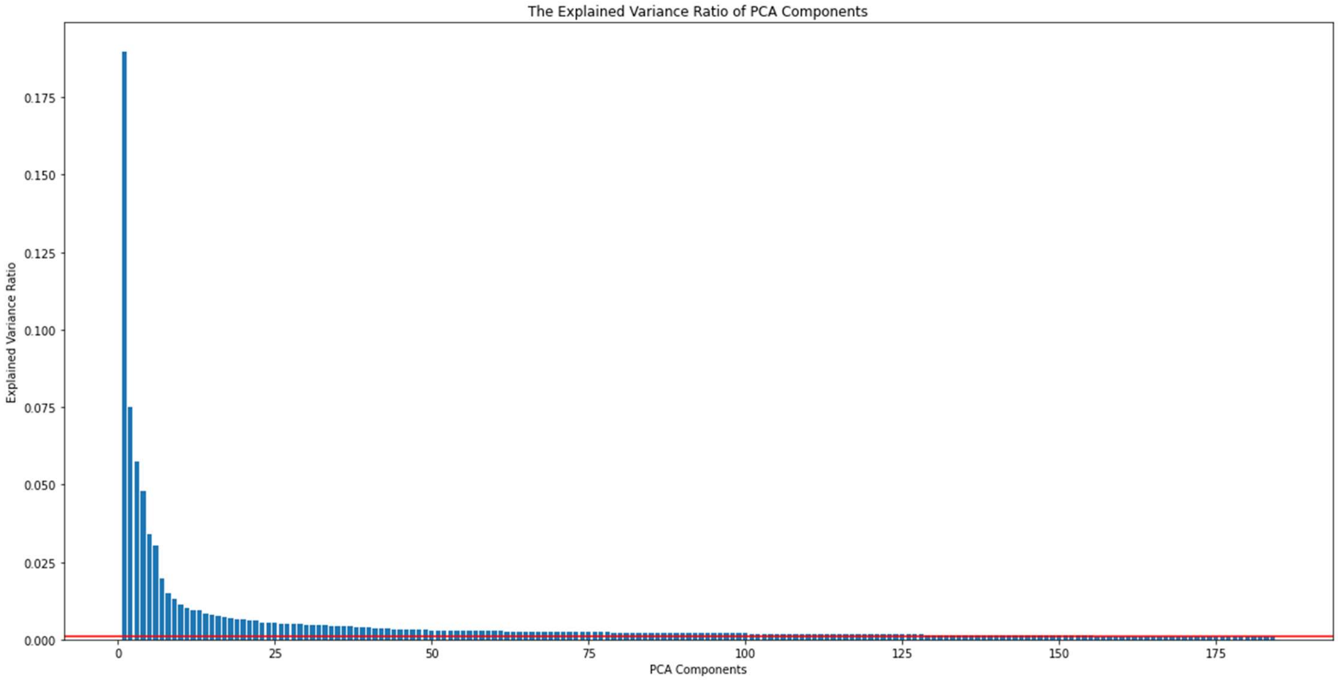
Implementation

Part I Unsupervised Modeling:

KMeans algorithm will be used for our problem as we need to segment the customers into clusters and find the common attributes between the customers in the same cluster.
And the data pipeline will go through the following process.

1- Data scaling: It's important to scale the data before fitting it to the model to get better predictions. Sciklearn standard scaler transformer is used for this purpose.

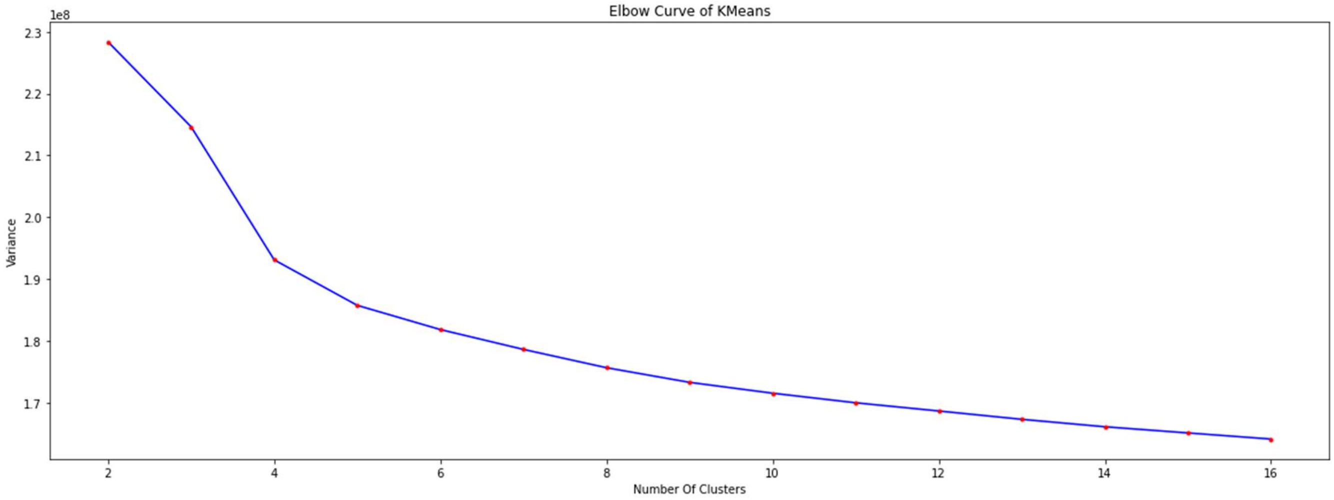
2- PCA is applied to the scaled data to reduce the dimensionality of the data set (364 columns).
The number of features is too high. Using PCA, we can measure the variance explained by PCA components. And it is shown in the graph below.



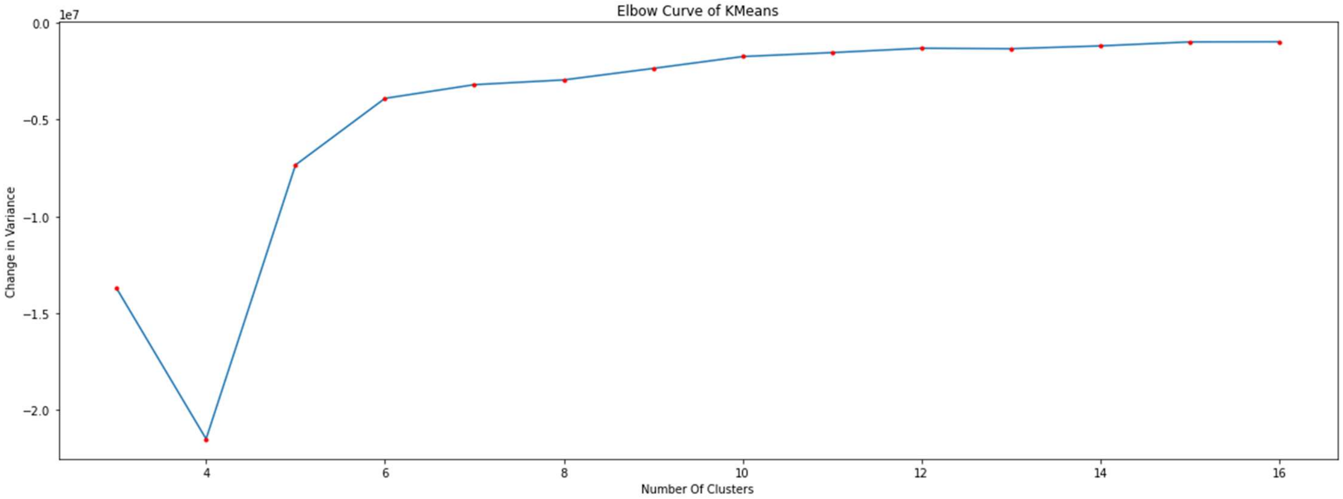
From the graph, we can see that half of the PCA components contain 93.6% of the variance explained by the components.

We can select only 182 components and use them in the following stages of the models.

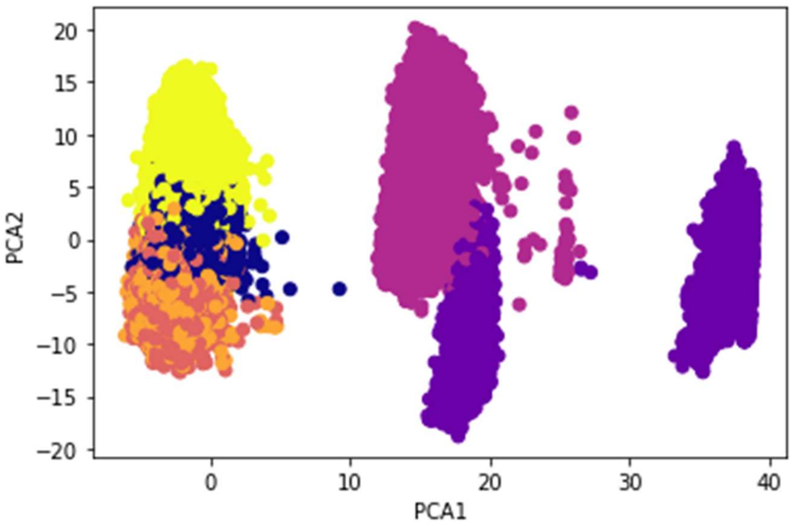
3- The KMeans model is fitted with the first 182 PCA components with a different number of clusters (from 2 clusters to 16 clusters). An elbow curve will be plotted.



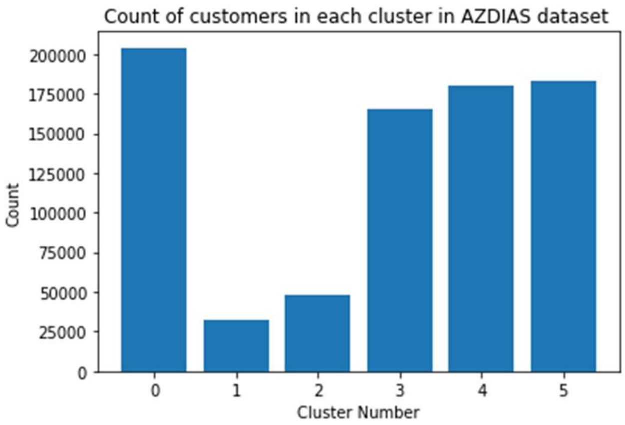
From the elbow curve, we see that 8 clusters are the break-out point that the total variance between the centroids and data points decreased. The change in variance curve of the elbow breakpoint can be shown more clearly.

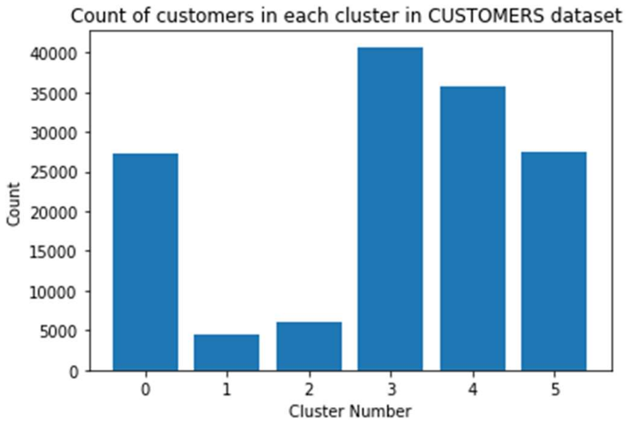


The next step is to use the KMeans model to predict the clusters of customers in AZDIAS and customers and find the common attributes of the segments. As illustration for the segments of the customers, the plot shown below between the first and second PCA components.

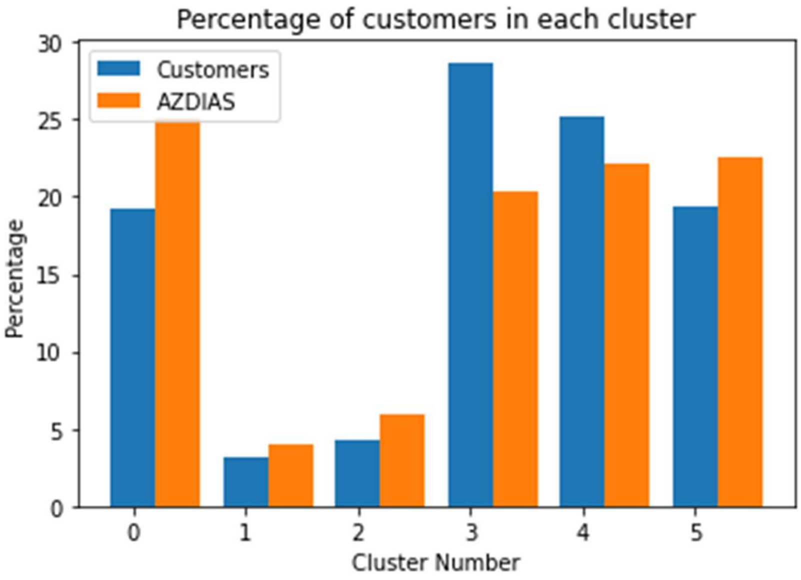


The count of customers in each segment in general population and customers data sets can be shown in the following plots.





By comparing the distribution of customers in different clusters, we see that they have roughly the same propotional distribution. This can be shown in the following plot.



From The distribution of the customers on the clusters, we can conclude the following points.

- Customers data set and AZDIAS data set have both roughly the same distribution of clusters.
- Majority of customers belong to cluster 3, however this cluster is not the major in the general population.
- Cluster 0 and cluster 5 has almost equal count in CUSTOMERS data set.
- We can study cluster 0 and cluster 5 as a group and cluster 3 and cluster 4 as one group.
- More focus on cluster 3 can lead to some of the common characteristics of the customers.

By grouping the customers by different clusters and determing the most frequent value in each feature with a threshold of 50%, we can see the common attributes of different clusters. By analysing the result data, we can conclude the following points.

Cluster 3 customers have the following common attributes.

- They have a very low financial minimalism behaviour and very high financial saving behaviour.
- They live in residential buildings.
- They have one active household in the cell.
- Their customer journey topology categories are, ‘advertising and consumption minimalist’ and ‘advertising and cross-channel ethaustism’
- Their share of car per household within PLZ8 is average.

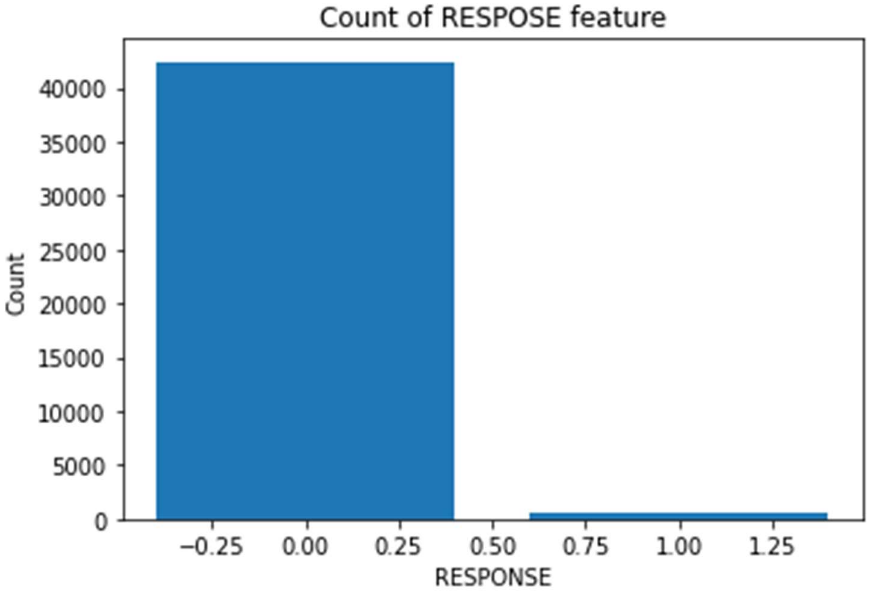
Cluster 0 and Cluster 5 customers have the common attributes.

- Most of them have german nationality as sounds from their names.
- The age of most of them is more than 60 years old.
- The share of small, upper class cars and van are average.
- Their return type is ‘determined Minimal-Returner’

We can see that cluster 3 customers belong to the middle-class with conservative consumption behavoiur, while customers who belong to cluster 0 and cluster 5 have more consuming habits. As a result, the company should focus on offers to cluster 3 customers (majority of the customers) that give more money saving.

Part II Supervised Modeling

In our use case, the independent variable (RESPONSE) is imbalanced as shown in the following plot.



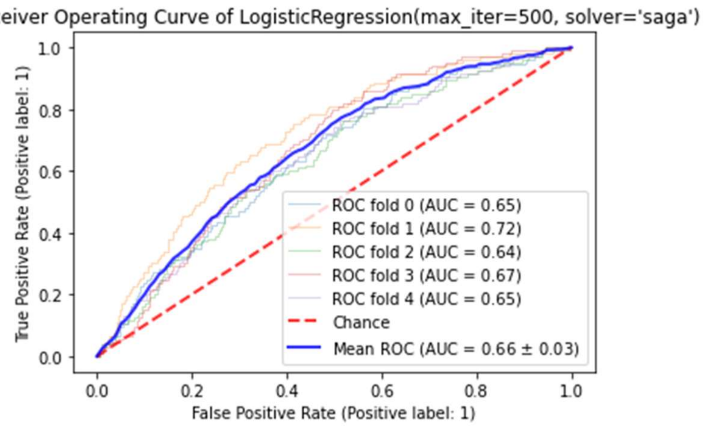
For this issue, stratified K-Fold Cross Validation is used to make sure that the folds have balanced distribution of both classes.
The following classification models will be tested to find the model with the best performance.

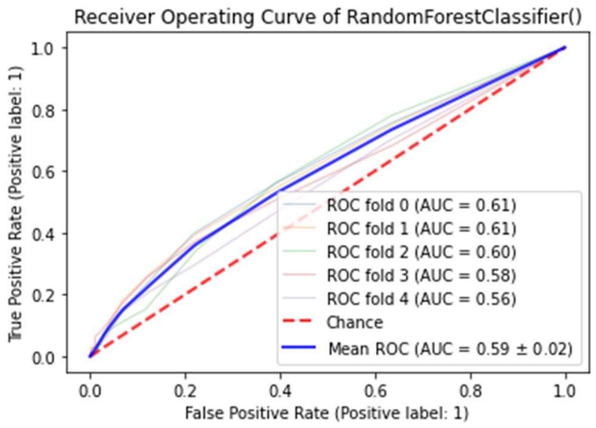
- Logistic Regression
- Random Forest
- Support Vector Machines
- Gradient Boosting Classifier

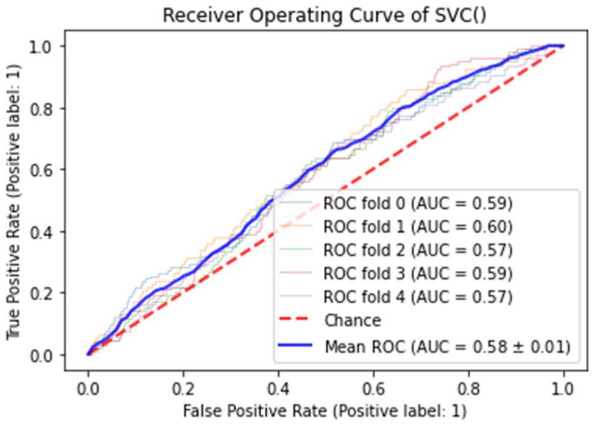
ROC curve is created for each model. The model with a higher AUC is the better model. The accuracy cannot be used as a metric in imbalanced datasets as the proportions of each class are not balanced. As a result, AUC can tell us how well the model is.

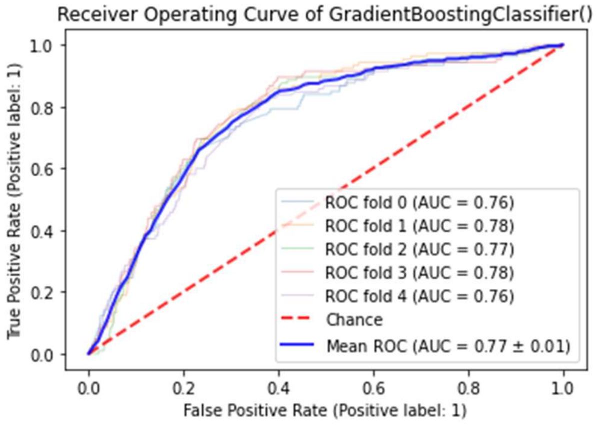
As a start point, only those four classification models have been selected and the default parameters were used except for Logistic regression (maximum iteration was raised to 500 because a warning raised that maximum iterations reached, and the solver changed to saga).

The default parameters for the model used to get an initial indication of the performance, and further investigations can be done for the targeted models.









From the resultant ROC curves, with 5 stratified K-Folds, it is seen that Gradient Boosting Classifier has the best mean AUC. Therefore, we will use the hyper-parameter tuning techniques to find the best performance model.

Refinement

For hyper-parameters, Bayesian Search optimization is chosen to be used as it takes less time than GridSearch optimization, especially with the large number of parameters to be tested. The search space used in hyper-parameters tuning has ranges of same of parameters of the model. More parameters can be added to the search space but with more computational and time resources.

The computational resources and time were a limitation factor in the refinement process. The GridSearch optimization goes through every possible combination of the parameters and this was expected to take too long (more than 9 days). Another approach should have been followed, and Bayesian optimization was a better choice as it goes through the promising combination of parameters rather than testing all parameters.

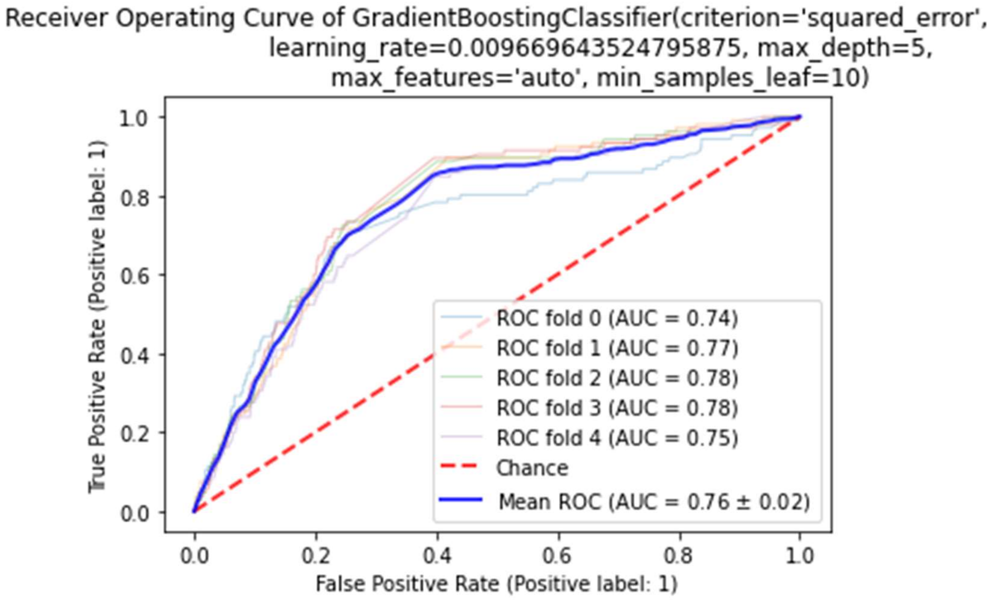
A quite limited search space was used for the Bayesian Search optimization as follows. The running time almost 8 hours with the following search space parameters. The Bayesian search optimizer saved too much time with compared to GridSearch.

```
'loss': ['deviance', 'exponential'],
'learning_rate': [1e-3,0.01],
'n_estimators': [int(x) for x in np.linspace(100,500,num=5)],
'criterion': ['friedman_mse', 'squared_error'],
'min_samples_split': [2,3,10,15],
'max_depth': [int(x) for x in np.linspace(5,20,num=4)],
'min_samples_leaf': [ 5, 10, 15, 20],
'max_features': ['auto','sqrt','log2']
```

As a result from Bayesian Search, the parameters of the best-performance model are as follows.

```
'criterion': 'squared_error',
'learning_rate': 0.009669643524795875,
'loss': 'deviance',
'max_depth': 5,
'max_features': 'auto',
'min_samples_leaf': 10,
'min_samples_split': 2,
'n_estimators': 100
```

From the ROC curve of the optimized model, we see that validation AUC is 0.768131344536 4253, that is the best result we can get from Bayesian Search.



Results

Model Evaluation and Validation

We see that there is not much improvement in the model with the hyper-parameters tuning , so a wider search space can be used to improve the score.
Now we can implement the optimized model to predict the response of the mail-out test dataset and by contributing in the Kaggle competition the score of the model with the predictions come from the model is 0.79589.

It's noted that in our case, we need a model with higher sensitivity because we need to capture any customer who could be interesting in the campaign and we don't care if we have many False Positives.

Justification

The score of testing data set is almost the same as the validation score which means that the model does not suffer from over-fitting problem.
A variety of classifications models can be tested to get higher score than can make the model deployed in production.

Conclusions

Customer segmentation with unsupervised ML models shows that the general population can be divided into 6 clusters.
Most of the Arvato customers belong to cluster 3 and they have the following characteristics.

- They have a very low financial minimalism behaviour and very high financial saving behaviour.
- They live in residential buildings.
- They have one active household in the cell.
- Their customer journey topology categories are, 'advertising and consumption minimalist' and 'advertising and cross-channel ethaustism'
- Their share of car per household within PLZ8 is average.

These aspects should be taken into consideration in any marketing campaign and here are the recommendations, the company should focus on offers and advertising campaigns that aim save money.

Customers who belong to cluster0 and cluster5 have a valuable weight to the overall customers' database which means, the company can design different marketing campaigns with different aspects like quality-based offers as the users in these clusters are not very financially conservative.

The best model to predict the response of target customers is GradientBoosting Classifier with mean AUC equals to 0.79589.

Reflection

The Cross-Industry Standard Process for Data Mining (CRISP-DM) process will be followed in two parts with the following steps.

1. Business understanding.
2. Data understanding.
3. Data preparation.
4. Modeling.
5. Evaluation.
6. Deployment.

An unsupervised modeling technique carried out to segment and classify the general population. This segmentation helped us in finding the common pattern for the target customers. KMeans algorithm used to classify the data points and PCA dimension reduction technique used.

A supervised model was implemented to predict the response of the customers to a mail marketing campaign. The best-performance model was Gradient Boosting classifier.

Improvement

The following points can be enhanced and are still open for more improvements.

- Many PCA components can be tested with an unsupervised ML model to find the number of components that leads to the best performance. This approach needs more computational resources.
- More classification models can be tested like XBoosting Classifier, LGBM and evaluate the performance in each case.
- The unsupervised and supervised models can be integrated into a web application that predicts the segment of any new customer and if the customer will respond to the marketing.

References

The data sets provided by Arvato Financial solution company under the terms and agreement of the data.