Capstone Project

Data Scientist Nanodegree

Avah Somtochukwu

June 20, 2023

Customer Segmentation Report – Arvato Financial Services

Definition

Project Overview

The demographics data for customers of a mail-order sales company in Germany is to be analyzed and compared against demographics information for the general population. We want use unsupervised learning techniques to perform customer segmentation so as to identify the parts of the population that best describe the core customer base of the company.
There are four data files associated with 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).
- 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).

The aim of this project is to build a model to predict which individuals are most likely to convert into becoming customers for the company.
The data has been provided by Udacity partners Bertelsmann Arvato Analytics

Problem Statement

The business problem we are trying to uncover is:

How can the mail-order sales company in Germany acquire new customers/clients more efficiently. Customer acquisition is vital so as to perform targeted marketing.

Metrics

The metrics used for this project are MAE, MSE, RMSE, r2-score and finally statsmodel. For MSE, MAE, RMSE, r2-score and statsmodel are: 2.382e-27, 3.65e-14, 4.88e-14, 1 and 1. respectively.

Analysis

Data Exploration and Visualization

There are 891221 fields and 366 attributes for Azdias while for customers there are 191652 fields and 369 attributes for Customers. There are no extra columns for Azdias dataset while for customers dataset there are extra columns. The extra columns are PRODUCT_GROUP, CUSTOMER_GROUP, and ONLINE_PURCHASE. Checked for missing values column-wise for each row and discovered that ALTER_KIND1, ALTER_KIND2, ALTER_KIND3, ALTER_KIND4 has 90%, 96%, 99%, 99% of null values. So, they were dropped. For other features with lesser null values, we just dropped the NAN value. For the other features with insignificant null values, performed a mean imputation whereby filled the NAN with the mean of the dataset. A function that handles these operations was created for the pre-processing of the Azdias and Customers dataset. For the extra columns in the customers dataset, the plot is as shown below:

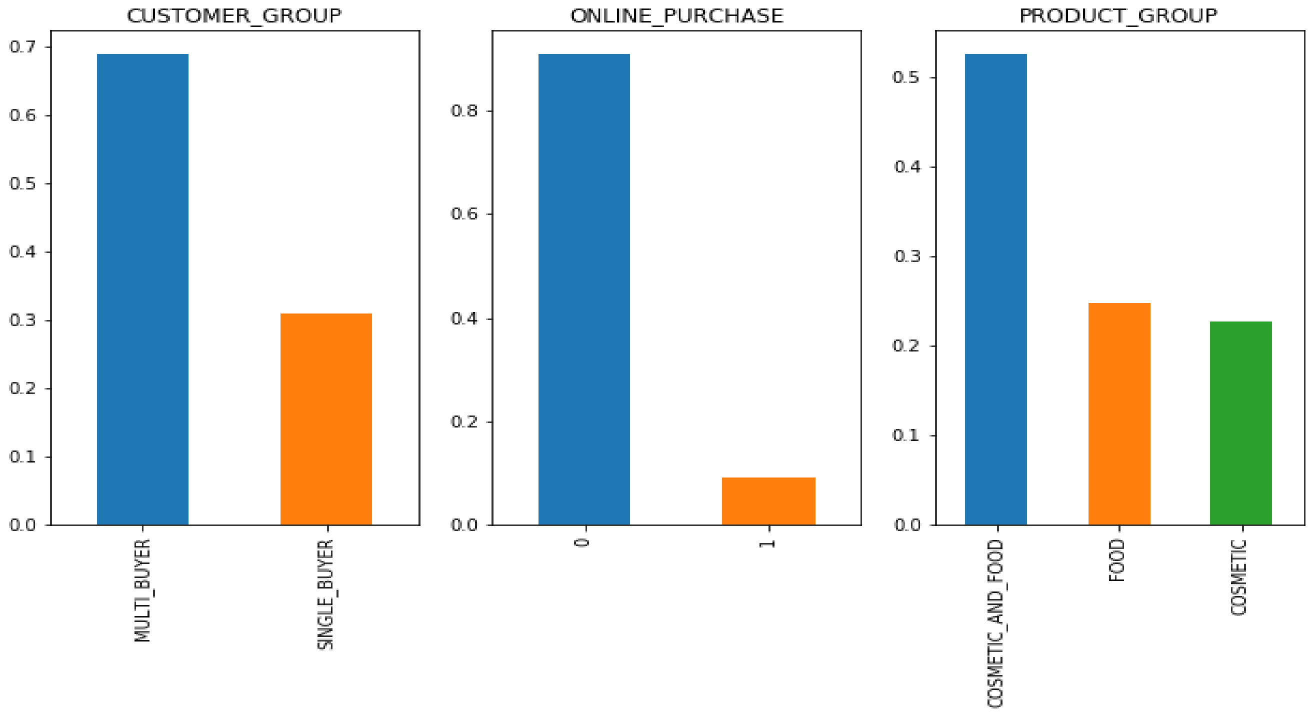


Fig 1. PRODUCT_GROUP, CUSTOMER_GROUP, and ONLINE_PURCHASE Graph

Plotted the distance plot of the LNR attribute of the Azdias and Customers dataset as shown below

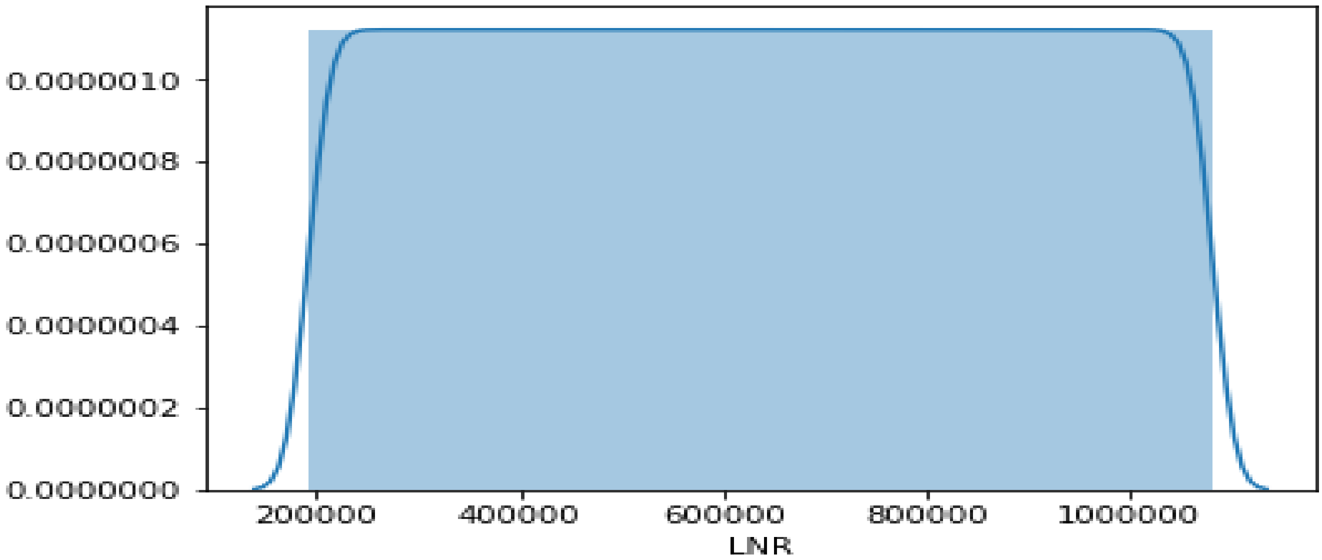


Fig 2. LNR attribute of Azdias distance plot

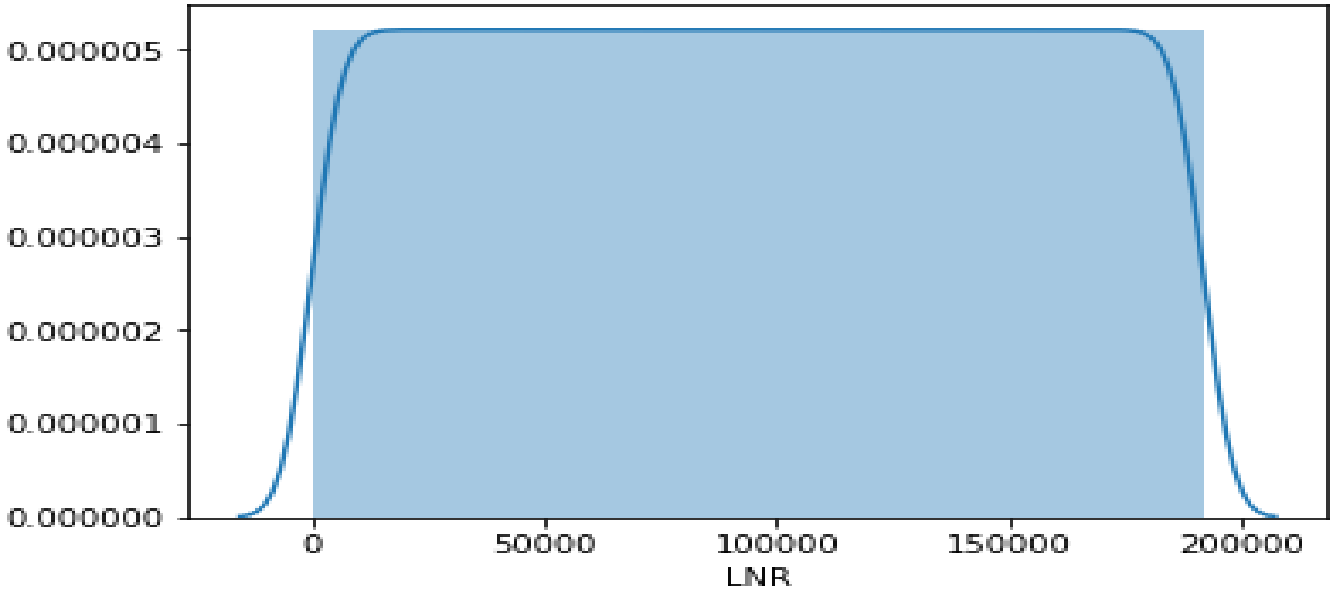


Fig 3. LNR attribute of Customers distance plot

Performed a Univariant analysis so as to get the descriptive statistics of the Azdias and the Customers dataset

Calculated the sum of the categorical values for the Azdias and the Customers dataset.

Algorithms and Techniques

1. Customer Segmentation

This section contains the main bulk of this project analysis. I applied unsupervised learning techniques to describe the relationship between the demographics of the company's existing customers and the general population of Germany. The unsupervised learning techniques applied are: K-Means, PCA and DBSCAN

K-Means

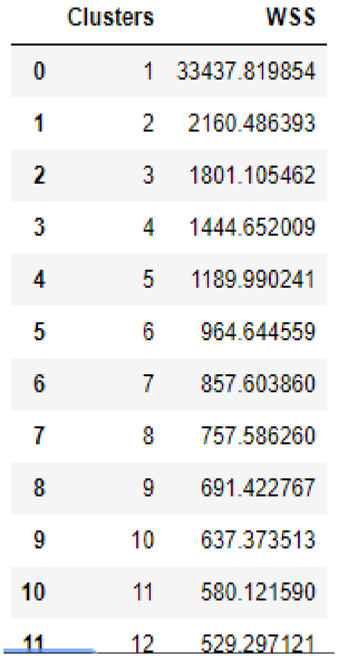
It involves grouping the unlabeled dataset into different clusters by looking for a fixed number (k) of clusters in the dataset. A cluster is a similar group of data in a large dataset.

A target number k=6 was defined which is simply the number of clusters, the maximum iteration was set to 200. The maximum iteration is the number of maximum iterations for each initialization of the k-means algorithm.

WCSS (Within Cluster Sum of Squares) is simply the sum of the squared distance between each member of the cluster and its centroid. The maximum iteration here is 300 with a random state of zero.

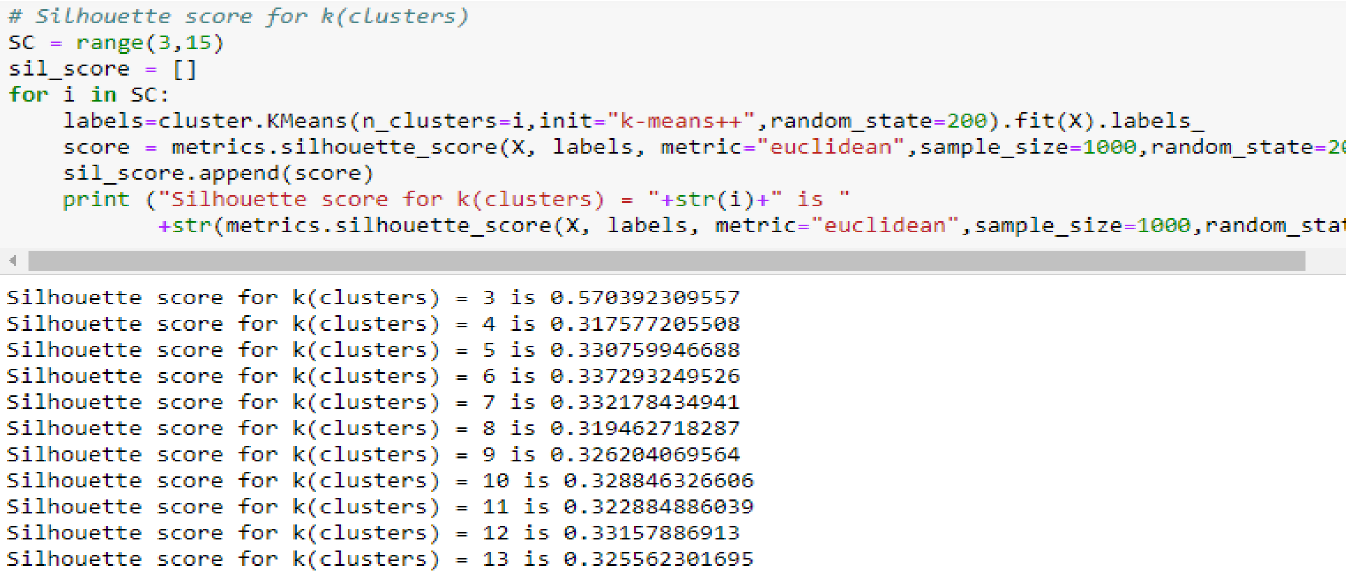
Elbow Method was used to determine the number of centroids(k) to use in the k-means.

Centroid is the multi-dimensional average of the cluster and is shown as a dataframe below:



	Clusters	WSS
0	1	33437.819854
1	2	2160.486393
2	3	1801.105462
3	4	1444.652009
4	5	1189.990241
5	6	964.644559
6	7	857.603860
7	8	757.586260
8	9	691.422767
9	10	637.373513
10	11	580.121590
11	12	529.297121

The Silhouette score is a representation of how well the data point has been clustered, scores above 0 are seen as good, while negative points mean the K-means algorithm has put that data point in the wrong cluster.



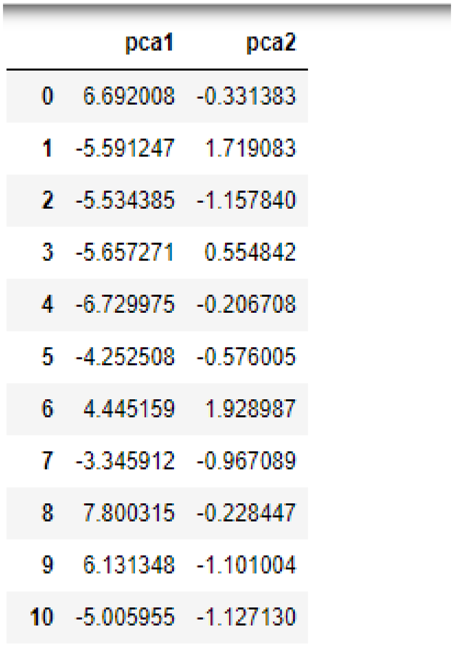
```
# Silhouette score for k(clusters)
SC = range(3,15)
sil_score = []
for i in SC:
    labels=cluster.KMeans(n_clusters=i,init="k-means++",random_state=200).fit(X).labels_
    score = metrics.silhouette_score(X, labels, metric="euclidean",sample_size=1000,random_state=2
    sil_score.append(score)
    print ("Silhouette score for k(clusters) = "+str(i)+" is "
           +str(metrics.silhouette_score(X, labels, metric="euclidean",sample_size=1000,random_stat
```

```
Silhouette score for k(clusters) = 3 is 0.570392309557
Silhouette score for k(clusters) = 4 is 0.317577205508
Silhouette score for k(clusters) = 5 is 0.330759946688
Silhouette score for k(clusters) = 6 is 0.337293249526
Silhouette score for k(clusters) = 7 is 0.332178434941
Silhouette score for k(clusters) = 8 is 0.319462718287
Silhouette score for k(clusters) = 9 is 0.326204069564
Silhouette score for k(clusters) = 10 is 0.328846326606
Silhouette score for k(clusters) = 11 is 0.322884886039
Silhouette score for k(clusters) = 12 is 0.33157886913
Silhouette score for k(clusters) = 13 is 0.325562301695
```

As seen from above the Silhouette score for the k(clusters) of 3-15 is well above zero and hence good.
Labels are algorithmically generated and assigned to data points in clusters. The label of each cluster is determined by examining the average characteristics of the observations classified to the cluster relative to the averages of those relative to the other clusters.
K-means cluster center is the arithmetic mean of all the points belonging to the cluster.

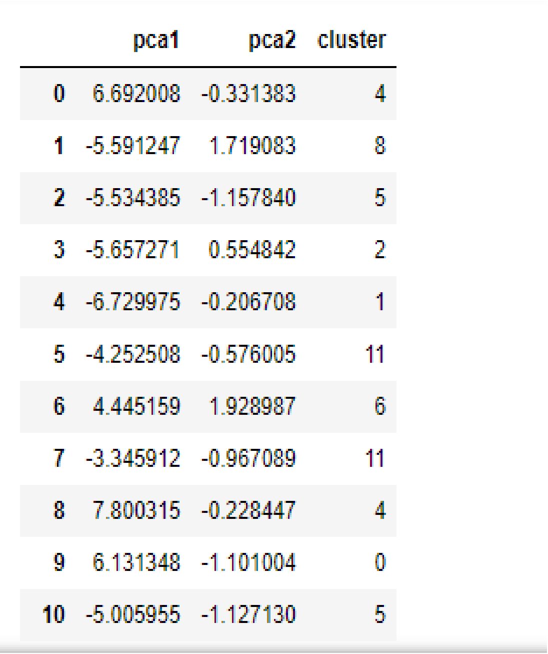
Principal component analysis (PCA)

A statistical approach that reduces the dimensions of a dataset by drawing strong patterns from the given dataset. The PCA dataframe is shown below.



	pca1	pca2
0	6.692008	-0.331383
1	-5.591247	1.719083
2	-5.534385	-1.157840
3	-5.657271	0.554842
4	-6.729975	-0.206708
5	-4.252508	-0.576005
6	4.445159	1.928987
7	-3.345912	-0.967089
8	7.800315	-0.228447
9	6.131348	-1.101004
10	-5.005955	-1.127130

Concatenated the pc dataframe with the dataframe of cluster with labels and is shown below:



	pca1	pca2	cluster
0	6.692008	-0.331383	4
1	-5.591247	1.719083	8
2	-5.534385	-1.157840	5
3	-5.657271	0.554842	2
4	-6.729975	-0.206708	1
5	-4.252508	-0.576005	11
6	4.445159	1.928987	6
7	-3.345912	-0.967089	11
8	7.800315	-0.228447	4
9	6.131348	-1.101004	0
10	-5.005955	-1.127130	5

The scatterplot of the two principal components is shown below:

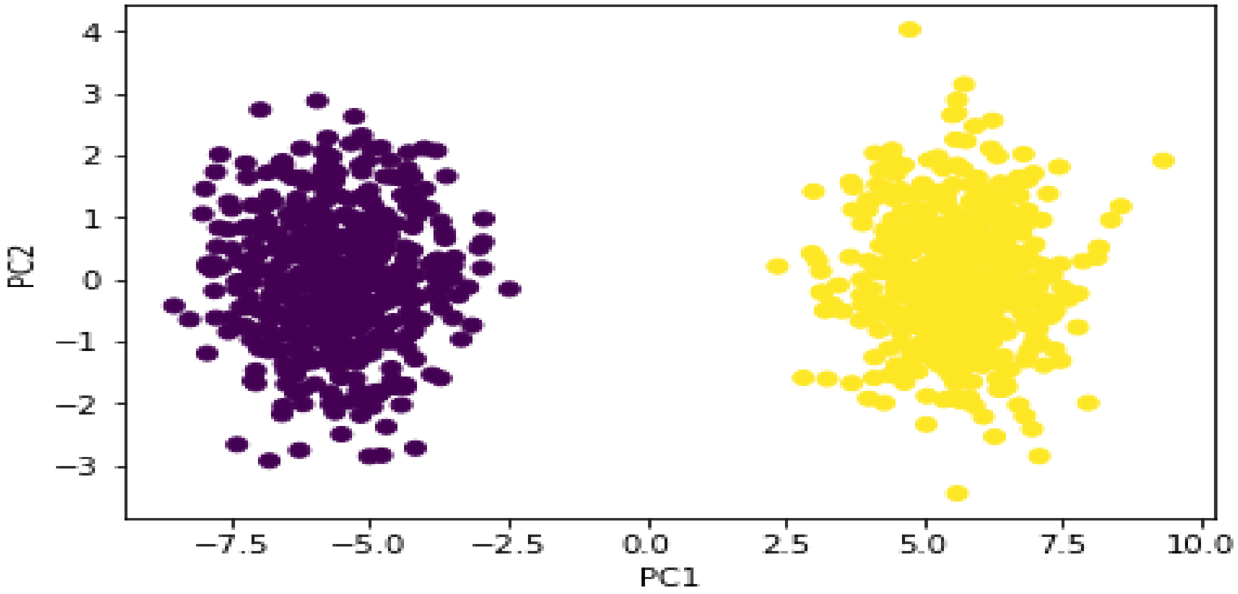


Fig 4. Scatterplot of PC1 and PC2

Density-based spatial clustering of applications with noise (DBSCAN)

Another statistical approach that finds arbitrary shaped clusters and clusters with outliers by locating a point that is close to many points from a cluster.

Eps specifies how close points should be to each other to be considered a part of a cluster.

Minimum sample is the minimum number of samples to form a dense region.

The scatterplot of Customers and Azdias datasets is shown below:

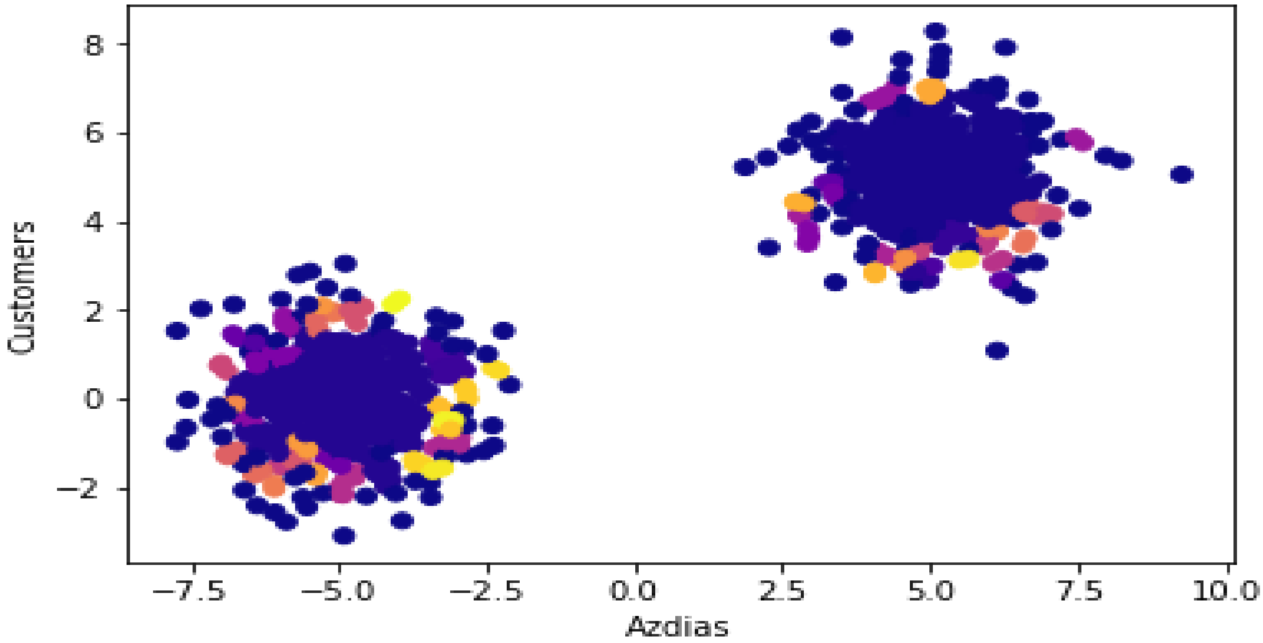
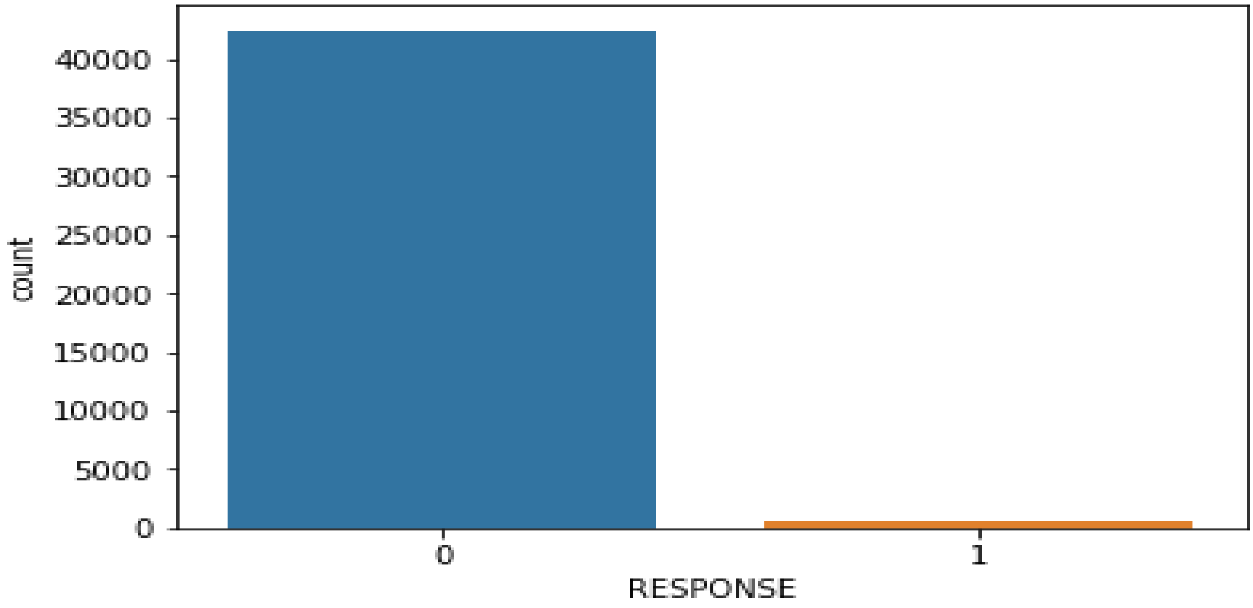


Fig 5: Scatterplot of Customers and Azdias datasets

2. Supervised Learning Model

Having found which parts of the population that are more likely to be customers of the mail-order company, we want to build a prediction model. Each of the rows in the "MAILOUT" data files represents an individual that was targeted for a mailout campaign. Ideally, we want to use the demographic information from each individual to decide whether or not it will be worth it to include that person in the campaign.

After printing the first five rows, we discovered that there is a column named “Response” hence we will use it to determine those that will respond to the email campaign. The Count plot is shown below:



Methodology

Data Pre-processing

Only three pre-processing was required. The first and second pre-processing was for Azdias and Customers. The final pre-processing was for the mail-out train dataset.

In the preprocessing of Azdias dataset, while checking for the number of unique values it was discovered that it contains duplicates. Missing values was checked for and columns with the highest number of null values i.e., at least 90% of null values were dropped. While for columns with second highest null values the NAs were dropped. Performed mean imputation for the rest of the columns with insignificant NAs. The same preprocessing holds for the customers dataset. For the mail-out-train pre-processing the same holds except the output returns a dataframe and a response dataframe.

Implementation

The different algorithms were implemented using unsupervised and supervised learning models.

Refinement

The model can be refined if more Customers and Azdias data was provided

Results

Model Evaluation and Validation

The model that turns out to be the best according to the score in test is

1. GradientBoostingClassifier() and
2. DecisionTreeClassifier()

Justification

The results obtained seems good enough. For more clarification. Kindly check out my Github

Conclusion

Reflection

Three datasets were provided whereas two was used for customer segmentation I.e., unsupervised learning models. The third dataset was used to build a prediction model from customers responses using supervised learning models.

The first two datasets were accessed and explored. During data exploration it was discovered that there were columns with 90% null values hence we dropped them. We also discovered columns with NAs we dropped the NAs. Then the rest of the columns we applied mean imputation. We also normalized/scaled down the data using the Minmax Scaler.

We checked for extra columns unfortunately the Azdias dataset has no extra columns but the Customers dataset has extra columns such as Product Group, Customer Group and Online Purchase. We visualized the data by plotting the extra columns.

Performed unsupervised learning techniques such as Kmeans, PCA and DBSCAN.

In the third dataset we applied supervised learning techniques such as Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier and Regressor. We discovered that Gradient Boosting Classify and Decision Tree Classifier provided better scores that the rest.

Improvement

The model can be further improved by using other algorithms and testing the model against those algorithms.