# Customer Segmentation Report for Arvato Financial Solutions

## 1. Project overview

Arvato Financial Solutions would like to use demographic information from individuals to decide whether or not it is worth it to include the individual in the campaign. The project is designed to address the challenge through unsupervised learning and supervised learning. The unsupervised learning techniques will be used to perform customer segmentation for a company for identifying the parts of the population that best describe the core customer base of the company, by exploring two dataset “Udacity_AZDIAS_052018.csv” and “Udacity_CUSTOMERS_052018.csv”. After that, supervised learning techniques will be used to make prediction on another two datasets “Udacity_MAILOUT_052018_TRAIN.csv” and “Udacity_MAILOUT_052018_TEST.csv”.

## 2. Data preprocessing

1) 3 data files were loaded

### Udacity_AZDIAS_052018

| | Unnamed: 0 | LNR | AGER_TYP | AKT_DAT_KL | ALTER_HH | ALTER_KIND1 | ALTER_KIND2 | ALTER_KIND3 | ALTER_KIND4 | ALTERSKATEGORIE_FEIN | ... | VHN | VK_DHT4A |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 910215 | -1 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN |
| 1 | 1 | 910220 | -1 | 9.0 | 0.0 | NaN | NaN | NaN | NaN | 21.0 | ... | 4.0 | 8.0 |
| 2 | 2 | 910225 | -1 | 9.0 | 17.0 | NaN | NaN | NaN | NaN | 17.0 | ... | 2.0 | 9.0 |
| 3 | 3 | 910226 | 2 | 1.0 | 13.0 | NaN | NaN | NaN | NaN | 13.0 | ... | 0.0 | 7.0 |
| 4 | 4 | 910241 | -1 | 1.0 | 20.0 | NaN | NaN | NaN | NaN | 14.0 | ... | 2.0 | 3.0 |

5 rows × 367 columns

### Udacity_CUSTOMERS_052018

| | Unnamed: 0 | LNR | AGER_TYP | AKT_DAT_KL | ALTER_HH | ALTER_KIND1 | ALTER_KIND2 | ALTER_KIND3 | ALTER_KIND4 | ALTERSKATEGORIE_FEIN | ... | VK_ZG11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 9626 | 2 | 1.0 | 10.0 | NaN | NaN | NaN | NaN | 10.0 | ... | 2.0 |
| 1 | 1 | 9628 | -1 | 9.0 | 11.0 | NaN | NaN | NaN | NaN | NaN | ... | 3.0 |
| 2 | 2 | 143872 | -1 | 1.0 | 6.0 | NaN | NaN | NaN | NaN | 0.0 | ... | 11.0 |
| 3 | 3 | 143873 | 1 | 1.0 | 8.0 | NaN | NaN | NaN | NaN | 8.0 | ... | 2.0 |
| 4 | 4 | 143874 | -1 | 1.0 | 20.0 | NaN | NaN | NaN | NaN | 14.0 | ... | 4.0 |

5 rows × 370 columns

### DIAS Attributes - Values 2017

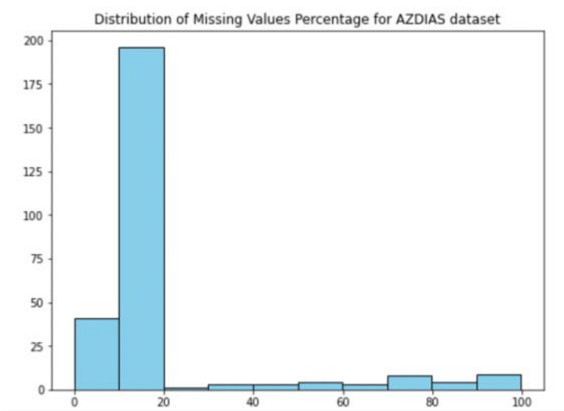
| | Attribute | Description | Value | Meaning |
|---|---|---|---|---|
| 0 | AGER_TYP | best-ager typology | -1 | unknown |
| 1 | NaN | NaN | 0 | no classification possible |
| 2 | NaN | NaN | 1 | passive elderly |
| 3 | NaN | NaN | 2 | cultural elderly |
| 4 | NaN | NaN | 3 | experience-driven elderly |

2) Identify missing value

The AZDIAS dataset has 367 columns and CUSTOMERS dataset has 370 columns, however, only 272 columns of both datasets can find description in Attributes dataset. So I decide to keep the 272 columns and remove the rest of columns from both datasets. According to the DIAS Attributes dataset, some attributes have meanings such as “unknown value” or “no classification possible”. These values are considered to be missing value and should be replaced with NA.

3) Remove columns with large portion of missing value

After above replace, percentage of missing value for each column for CUSTOMERS and AZDIAS datasets is calculated and plot into a histogram, as below. According to the charts, for AZDIAS dataset, most columns have missing value which are less than 20% while for CUSTOMERS dataset, most columns have missing value which are less than 40%. Then I decide to remove columns with more than 20% missing value for AZDIAS dataset and remove columns with more than 40% missing value for CUSTOMERS dataset.





4) Check columns with object data type

Columns with object data type can hold various types of data. 3 columns of CUSTOMERS dataset and AZDIAS dataset are found with mixed datatype, include: “CAMEO_DEUG_2015”, “CAMEO_DEU_2015”, “OST_WEST_KZ”. After checking the attribute dataset with all possible values under the 3 columns, I decide that all the 3 columns represent categorical variables, and their value should be string format. So all the 3 columns are transformed to string format. After that, all these 3 columns (categorical variable) are transformed into numeric format using label encoding.

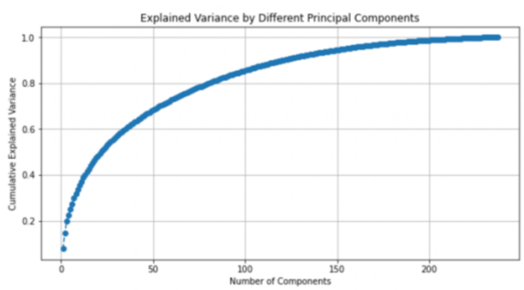
5) Impute missing value

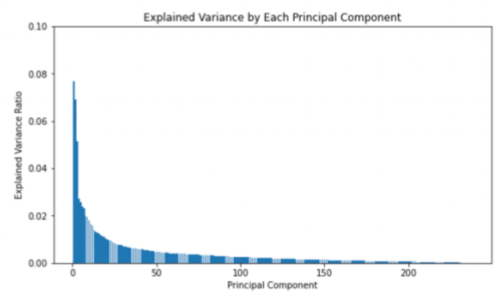
Forward fill method is used to impute the missing value for all columns of CUSTOMERS dataset and AZDIAS dataset. However, if there is no previous value in the column, the missing value remains NA. In this case, all the rest of missing value is imputed with the most frequent value in the column.

3. Customer segmentation

    1) Principal component analysis

Principal component analysis is a great technique to reduce dimensionality of large dataset. Before perform clustering, I use principal component analysis to reduce the dimensionality. Below two chart shows “Explained Variance by Different Principal Components” and “Explained Variance by Each Principal Component” after performing PCA on the CUSTOMERS datasets. As shows in the chart, when the number of components increases to 100, more than 80% variance can be explained. I decide to use 100 components to perform clustering in the later part.




Before diving into clustering, I would like to look at the first three components individually, which explain the variance most.

1st component

As showed below, the characteristic of this group is related to the car owned by the person, such as share of luxury cars, share of cars with high max speed or share of small and very small cars (Ford Fiesta, Ford Ka etc.).

| | Attribute | Description |
|---|---|---|
| 203 | KBA13_HERST_BMW_BENZ | share of BMW & Mercedes Benz within the PLZ8 |
| 212 | KBA13_KMH_211 | share of cars with a greater max speed than 210 km/h within the PLZ8 |
| 213 | KBA13_KMH_250 | share of cars with max speed between 210 and 250 km/h within the PLZ8 |
| 236 | KBA13_MERCEDES | share of MERCEDES within the PLZ8 |
| 250 | KBA13_SEG_OBEREMITTELKLASSE | share of upper middle class cars and upper class cars (BMW5er, BMW7er etc.) |

| | Attribute | Description |
|---|---|---|
| 205 | KBA13_HERST_FORD_OPEL | share of Ford & Opel/Vauxhall within the PLZ8 |
| 209 | KBA13_KMH_180 | share of cars with max speed between 110 km/h and 180km/h within the PLZ8 |
| 211 | KBA13_KMH_140_210 | share of cars with max speed between 140 and 210 km/h within the PLZ8 |
| 227 | KBA13_KW_0_60 | share of cars up to 60 KW engine power - PLZ8 |
| 245 | KBA13_SEG_KLEINWAGEN | share of small and very small cars (Ford Fiesta, Ford Ka etc.) in the PLZ8 |

2nd component

The characteristic of this group appears to be related to socioeconomic profile such as financial typology, life stage, and social status.

| | Attribute | Description |
|---|---|---|
| 11 | CAMEO_DEUG_2015 | CAMEO classification 2015 - Uppergroup |
| 12 | CAMEO_DEU_2015 | CAMEO classification 2015 - detailled classification |
| 81 | FINANZ_SPARER | financial typology: money saver |
| 295 | SEMIO_KAEM | affinity indicating in what way the person is of a fightfull attitude |
| 296 | SEMIO_KRIT | affinity indicating in what way the person is critical minded |

| | Attribute | Description |
|---|---|---|
| 80 | FINANZ_MINIMALIST | financial typology: low financial interest |
| 271 | LP_LEBENSPHASE_FEIN | lifestage fine |
| 272 | LP_LEBENSPHASE_GROB | lifestage rough |
| 273 | LP_STATUS_FEIN | social status fine |
| 274 | LP_STATUS_GROB | social status rough |

3rd component

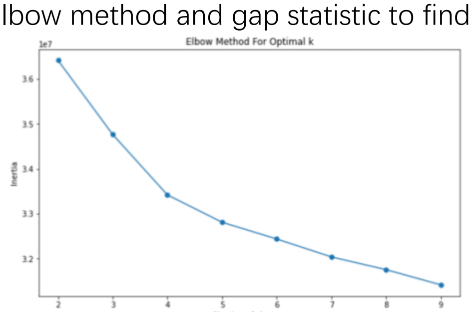
Like the characteristic of 2nd component, the characteristic of this group is also related to socioeconomic profile such as age, financial typology, and affinity of social minded.

| | Attribute | Description |
|---|---|---|
| 2 | ALTERSKATEGORIE_GROB | age classification through prename analysis |
| 83 | FINANZ_VORSORGER | financial typology: be prepared |
| 92 | HH_EINKOMMEN_SCORE | estimated household net income |
| 303 | SEMIO_SOZ | affinity indicating in what way the person is social minded |
| 305 | SEMIO_VERT | affinity indicating in what way the person is dreamily |

| | Attribute | Description |
|---|---|---|
| 78 | FINANZ_ANLEGER | financial typology: investor |
| 81 | FINANZ_SPARER | financial typology: money saver |
| 82 | FINANZ_UNAUFFAELLIGER | financial typology: unremarkable |
| 280 | OST_WEST_KZ | flag indicating the former GDR/FRG |
| 302 | SEMIO_REL | affinity indicating in what way the person is religious |

2) Clustering

The first 100 of PCA components were selected for performing clustering (K means). Then I use elbow method and gap statistic to find the optimal number of clusters, which is 9.



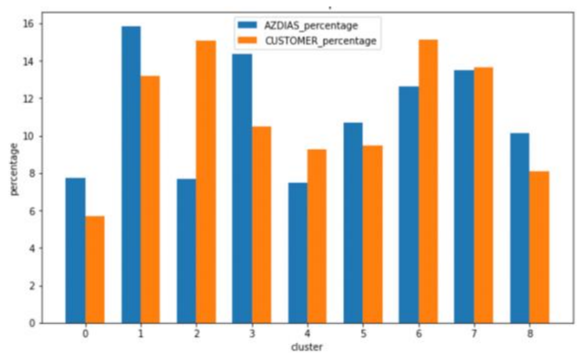

```
optimalK = OptimalK(parallel_backend='joblib')

# Calculating the optimal number of clusters
n_clusters = optimalK(CUSTOMER_pca, cluster_array=np.arange(2, 10))

print('Optimal number of clusters:', n_clusters)
Optimal number of clusters: 9
```

Once confirmed the number of clusters, I perform clustering for the AZDIAS dataset (demographics data for the population of Germany) and CUSTOMERS dataset (demographics data for customers of the mail order company. After that, the 9 clusters were mapped to the AZDIAS dataset and CUSTOMERS dataset and then we can see which cluster an individual is located in.

Below chart compare the population of Germany and customers of the mail order company by showing how much portion of individuals of the total population is in each of 9 clusters.




| | cluster | AZDIAS_percentage | CUSTOMER_percentage |
|---|---|---|---|
| **0** | 0 | 7.726927 | 5.710350 |
| **1** | 1 | 15.825480 | 13.168660 |
| **2** | 2 | 7.684850 | 15.046543 |
| **3** | 3 | 14.333931 | 10.508630 |
| **4** | 4 | 7.497691 | 9.238620 |
| **5** | 5 | 10.683658 | 9.451506 |
| **6** | 6 | 12.648378 | 15.122201 |
| **7** | 7 | 13.466132 | 13.641913 |
| **8** | 8 | 10.132952 | 8.111577 |

4. Supervised learning model

The supervised learning model is trained on a separate dataset "Udacity_MAILOUT_052018_TRAIN". Preprocessing techniques were performed on the dataset and then I use 3 machine learning models (random forest, logistic regression, gradient boosting) to perform the supervised training. As showed below, all the 3 models have the same best cross validation score. Finally, I chose the logistic regression model to make the prediction.

```
# Random forest
rf_model = RandomForestClassifier()
rf_params = {
    'n_estimators': [10, 50, 100],
    'max_depth': [None, 10, 20, 30]
}
rf_grid_search = GridSearchCV(rf_model, rf_params, cv=5, return_train_score=False)
rf_grid_search.fit(X_train, y_train)
print(f"Best parameters for RandomForestClassifier: {rf_grid_search.best_params_}")
print(f"Best cross-validation score for RandomForestClassifier: {rf_grid_search.best_score_}\n")
```

```
Best parameters for RandomForestClassifier: {'max_depth': None, 'n_estimators': 50}
Best cross-validation score for RandomForestClassifier: 0.987616967038911
```

```
# Logistic Regression
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
lr_model = LogisticRegression(max_iter=1000)
lr_params = {
    'C': [1, 5, 10]
}
lr_grid_search = GridSearchCV(lr_model, lr_params, cv=5, return_train_score=False)
lr_grid_search.fit(X_train_scaled, y_train)
print(f"Best parameters for LogisticRegression: {lr_grid_search.best_params_}")
print(f"Best cross-validation score for LogisticRegression: {lr_grid_search.best_score_}\n")
```

```
Best parameters for LogisticRegression: {'C': 1}
Best cross-validation score for LogisticRegression: 0.987616967038911
```

```
# Gradient Boosting
gbc_model = GradientBoostingClassifier()
gbc_params = {
    'n_estimators': [50, 100, 200],
    'learning_rate': [0.01, 0.1, 0.2]
}
gbc_grid_search = GridSearchCV(gbc_model, gbc_params, cv=5, return_train_score=False)
gbc_grid_search.fit(X_train, y_train)
print(f"Best parameters for GradientBoostingClassifier: {gbc_grid_search.best_params_}")
print(f"Best cross-validation score for GradientBoostingClassifier: {gbc_grid_search.best_score_}\n")
```

```
Best parameters for GradientBoostingClassifier: {'learning_rate': 0.01, 'n_estimators': 50}
Best cross-validation score for GradientBoostingClassifier: 0.987616967038911
```

5. Conclusion

In this project, both unsupervised and supervised learning methodologies are explored to refine the understanding of Arvato Financial Services' customer base. Through unsupervised learning, I conduct customer segmentation analysis, identifying distinct groups within the population that align closely with the company's primary clientele. Utilizing supervised learning techniques—specifically random forest, logistic regression, and gradient boosting—I develop predictive models to forecast customer behavior. The combined insights gained from our segmentation analysis and predictive modeling provide valuable intelligence that will inform and enhance Arvato's marketing strategies, ensuring they are targeted and efficient. However, given the constraints of time and resources, there are opportunities to further enhance the project's outcomes in future iterations, include improving the interpretation for the PCA results as well as the clusters results, developing ensemble models to make predictions and utilizing domain knowledge expert insights to understand which features may be more relevant.