

**Project:** Customer Segmentation and Identifying Potential Customers for Arvato Financial Services

**Problem statements:** A mail-order sales company in Germany would like to identify the parts of the Germany population that best describe the core customer base of the company. The company would then like to use the insights obtained from the customer segmentation to help developing models that can predict which individuals are more likely to convert to their customers so that they can perform more targeted and successful market campaign in the future.

**Domain Background:** This project is provided by Bertelsmann Arvato Analytics, which is an internationally active services company that develops and implements innovative solutions for business customers from around the world. These include Supply Chain Management solutions, financial services and IT services, which are continuously developed with a focus on innovations in automation and data/analytics.

This project lies in the growing field of customer analytics. Consumer product focused companies start to use internal data (such sales/transaction data, customer information) together with external data (such as population demographics, industry information) to better understand customers and their consumption behaviors. The insights from customer analytics can help companies improve business performance by conducting more targeted marketing to acquire potential customers as well as developing products and services that can better serve their existing customers and thus increasing sales.

I choose this project because I'm interested in the consumer and marketing analytics space and would like to improve my skills in this area.

**Datasets** are provided by Bertelsmann Arvato Analytics. There are four datasets in total.

- 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).

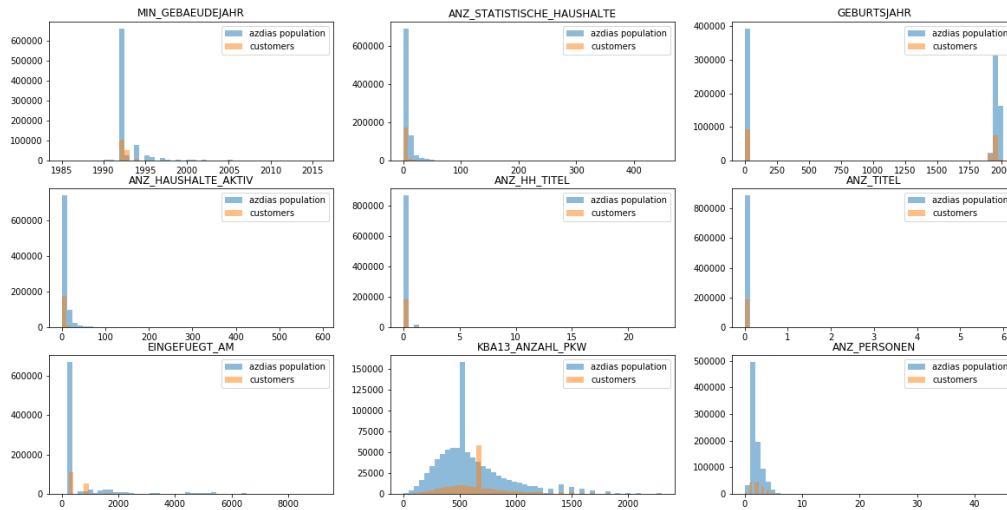
Each row of the demographics files represents a single person, but also includes information outside of individuals, including information about their household, building, and neighborhood. The "CUSTOMERS" file contains three extra columns ('CUSTOMER\_GROUP', 'ONLINE\_PURCHASE', and 'PRODUCT\_GROUP'), which provide broad information about the customers depicted in the file. The "MAILOUT TRAIN" file includes one additional column, "RESPONSE", which indicated whether or not each recipient became a customer of the company. All the insights learned from the first three data sets will be used to predict whether individuals in "MAILOUT TEST" set will respond or not.

Two Excel spreadsheets are provided.

- “DIAS Information Levels - Attributes 2017.xlsx” is a top-level list of attributes and descriptions, organized by informational category.
- “DIAS Attributes - Values 2017.xlsx” is a detailed mapping of data values for each feature in alphabetical order.

## Data Exploration

1. It is found that not all variables in the data sets are described in the attribute description files.
2. According to the attribute and value files, there are only 7 numeric variables:  
 'ANZ\_HAUSHALTE\_AKTIV', 'ANZ\_HH\_TITEL', 'ANZ\_PERSONEN', 'ANZ\_TITEL',  
 'GEBURTSJAHR', 'KBA13\_ANZAHL\_PKW', 'MIN\_GEBAEUDEJAHR'
3. I checked the number of unique values for each variable in order to identify additional potential numeric variables that are not covered by the attribute description files. It is important to understand whether a variable is intended as numeric or category, because categorical variables, even coded as numbers, should be encoded to dummy variables.  
 There are three variables with more than 100 unique values in azdias data.
  - LNR: Individual number
  - EINGEFUEGT\_AM: time. This will be changed to number of days since the earliest time in azdias
  - ANZ\_STATISTISCHE\_HAUSHALTE
4. Distribution of the numerical variables



5. When downloading the data, it is noticed that two variables, CAMEO\_DEUG\_2015 and CAMEO\_INTL\_2015, have values of mixed types (string and numeric).

- The description of CAMEO\_DEUG\_2015 is CAMEO classification 2015 – Upper group, which suggests each value represents a group. So converting all numbers to strings to be consistent with the supposed values in the values file.
- CAMEO\_INTL\_2015 doesn't exist in the attribute and value description files. However, would expect it to be similar to CAMEO\_DEUG\_2015. Therefore, convert everything to strings too.

6. Check proportion of missing values for each variable in customer data:

- Remove the variables which have more than 30% missing values
- For the variables that we keep

- If numeric, fill missing using mean
- If categorical, doesn't need to fill since we will create dummy variable for each group and can automatically interpret missing as one group

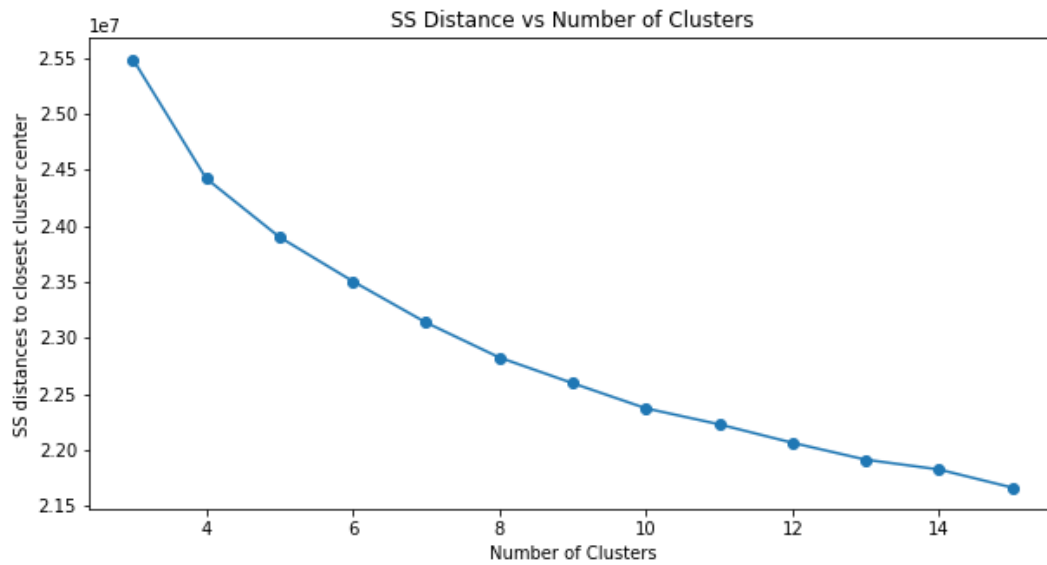
#### 7. Create dummy variables for all categorical variables

- Number of candidate categorical variable to be examined: 350
- For each candidate variable, only keep the dummy variable (group indicator) if
  - the mean of that group is statistically different between customers and azdias population (p-value 1%)
  - and the relative difference is larger than 50%
  - and at least 5% of people are in the group (for both customers and azdias population)
- After the filtering, 160 category variables are kept in the data, with 322 dummy variables in total

#### 8. Data is normalized using MinMaxScaler

#### 9. Clustering to understand customer segments

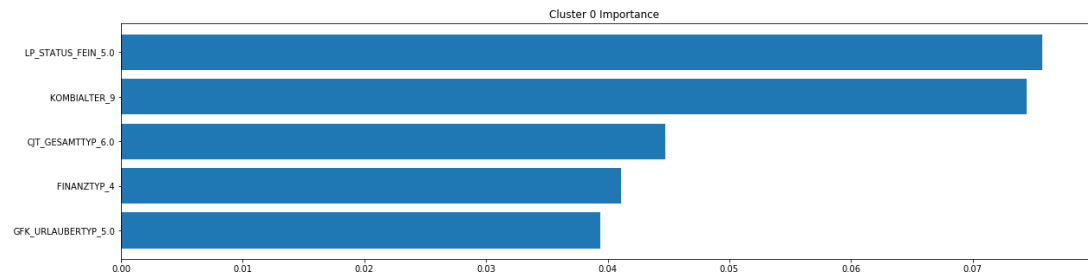
- Train KMeans clustering using azdias population data
  - Choose the number of clusters using elbow method : 10 clusters



- Obtain predicted clusters for azdias population and customers
- Compare the proportions in different cluster (customers vs population)
- Findings: Proportion of customers in cluster 0 and 8 are much larger than proportion of population in cluster 0 and 8. Therefore, for market campaign, targeting people in cluster 0 and 8 are more likely to succeed.

Cluster	0	8	4	7	9	6	1	2	5	3
Proportion of Customers	21.4%	44.3%	15.7%	8.0%	3.2%	2.3%	2.3%	1.5%	0.9%	0.5%
Proportion of Population	5.1%	15.7%	15.6%	9.4%	5.3%	7.2%	11.0%	9.5%	12.7%	8.4%
Ratio	4.2	2.8	1.0	0.9	0.6	0.3	0.2	0.2	0.1	0.1

- Example feature importance graph for cluster 0










- Particular profiles for cluster 0: minimalistic high-income learner, financially prepared, has some online affinity, Advertising-Enthusiast with restricted cross-channel-behavior, nature fans

#### 10. Prediction: trying to predict who may respond to market campaign

- Benchmark model: Naïve Bayesian Classifier
- Other Algorithms: Random forest classifier and XGBoost and ensemble
- Evaluation metrics: area under ROC curve.
- I first split the train data further to train (70%) and test (30%) set so that I can test the results before submitting results to Kaggle
- To choose the hyperparameters of each model, I use cross validation
  - For Random forest classifier
    - `n_estimators = 1500`
    - `max_depth = 6`
  - For XGBoost classifier
    - `learning_rate=0.05`
    - `max_depth = 3`

- | Model          | Area under ROC curve |
|----------------|----------------------|
| Naïve Bayesian | 0.605                |
| Random Forest  | 0.756                |
| XGBoost        | 0.760                |
| Ensemble       | 0.774                |

- | Overview | Data                     | Code     | Discussion  | Leaderboard | Datasets | ...  | My Submissions | Submit Predictions |
|----------|--------------------------|----------|---|-------------|----------|------|----------------|--------------------|
| #        | Team Name                | Notebook | Team Members  | Score ?     | Entries  | Last |                |                    |
| 1        | Oliver Farren            |          |  | 0.84739     | 5        | 5mo  |                |                    |
| 2        | Chin Danny               |          |  | 0.84521     | 77       | 1h   |                |                    |
| 3        | Ambresh Patil            |          |  | 0.81063     | 58       | 1y   |                |                    |
| 4        | Julio Guijarro Hernandez |          |  | 0.80954     | 16       | 9mo  |                |                    |
| 5        | [Deleted]                |          |  | 0.80954     | 12       | 9mo  |                |                    |
| 6        | Telmo                    |          |  | 0.80936     | 57       | 1y   |                |                    |
| 7        | Mia Liang                |          |  | 0.80907     | 6        | 3m   |                |                    |