Customer Segmentation Report Arvato Financial Solutions

Udacity Data Science Nanodegree Capstone Project

Camilo Mesa

Introduction

Demographics data for customers of a mail-order sales company in Germany is analyzed and compared against demographics information for the general population to improve customer acquisition. The data was provided by Bertelsmann Arvato Analytics, a company that offers financial services, Information Technology (IT) services, and Supply Chain Management (SCM) solutions for business customers on a global scale. Two main questions motivate this work:

- How can our client (a mail-order company) acquire clients more efficiently?
- Which people in Germany are more likely to become new customers of our client?

Using unsupervised learning, we present a customer segmentation of the company that identifies parts of the population that best describe its core customer base.

The information used in the customer segmentation is then incorporated in a model designed for targeting a marketing campaign more efficiently. We develop a classification model to predict which individuals are most likely to become new customers.

This project consisted of creating a data pipeline in which we mine the datasets as follows.

1. Data is cleaned and processed.
2. The most important features are selected using a customer classification model and their contribution to the prediction.
3. Principal components analysis is implemented to further reduce the data set's dimension and select a combination of the features that explained most of the data set variance.
4. The general population is segmented using clustering (k-means)**.** Then we used the distribution of customers among the clusters to find distinctive features of the company's typical clients. This analysis also helped engineered a feature used in the next step.
5. Finally, we trained a gradient boost classification algorithm to classify clients who were likely to respond to a mailout campaign. The final model had an area under the receiver operating characteristic curve score of 0.77.

The details of the computations and processes described in this report are available on this GitHub repository.

The Data

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).

Each row of the demographics files represents a single person and 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 was retained, but in the "TEST" subset, it was removed; our final predictions are assessed in a Kaggle competition against that withheld column.

Otherwise, all of the remaining columns are the same between the three data files. More information about the columns depicted in the files is available in the following two Excel spreadsheets:

- 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 Cleaning and Processing

The AZDIAS and CUSTOMERS data sets have 88246 common rows. A possible explanation for this is that rows in the CUSTOMERS data set correspond to clients who live outside of Germany.

Of the 366 demographic features, 51 were not described in the DIAS Information Levels - Attributes 2018.xlsx. We dropped these features from the datasets as interpretability of the demographics is one of our main goals.

We substituted values in columns with no descriptions for np.NaN and discarded sparse columns and rows. Most columns in AZDIAS have less than 20% of missing values, but a large number of columns in CUSTOMERS had 20-25% NaN.

We convereted to dummy variables features like ‘D19_LETZTER_KAUF_BRANCHE', 'OST_WEST_KZ’ with nonnumerical categories. Lastly, we used a mean imputer to fill in NaN values in every column.

The function ‘clean_data’ available in the python module under the same name contains all the cleaning and processing steps above.

Feature Selection and Dimensionality Reduction

Customer Classification Model and Gini Importance

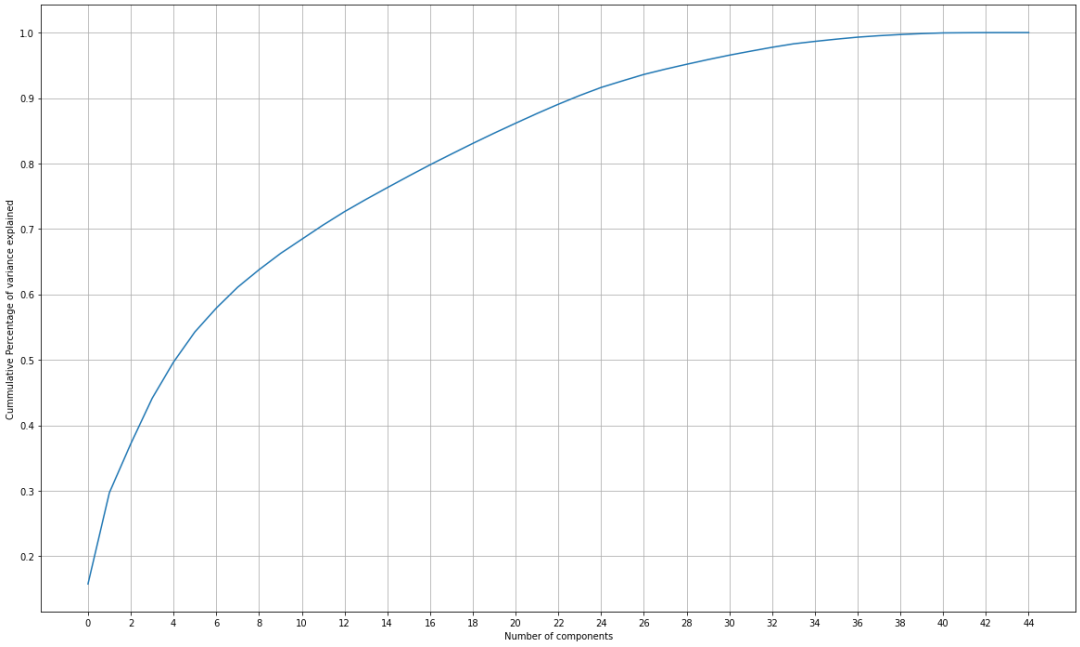
To select important features and begin our dimensionality reduction, we set up two classification models (AdaBoost and GradientBoostClassifier) to predict if a given person is a customer of the company. We merged the AZDIAS and CUSTOMERS data sets and added a column ‘CUSTOMER’ with a zero value if the row was from AZDIAS and not CUSTOMERS 1 else.

Then we selected features based on the Gini importance of both models. The Gini importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature. The higher the value, the more important the feature. We selected those features that had an accumulated sum of Gini importance of 0.99.

With this selection criteria, a total of 45 of the 280 features we had from the cleaning and processing were selected. This significant reduction proved to be important in making the customer segmentation more manageable and the acquisition process more lean and efficient.

Principal Component Analysis (PCA)

After using a StandarScaler to normalize both data sets, we proceeded to implement PCA to reduce the number of features and use the features that influence the variance of the general population data set. We decided to use the first 22 principal components, which explain almost 90% of the AZDIAS dataset variance. The following plot describes the cumulative percentage of the variance as a function of the number of principal components.

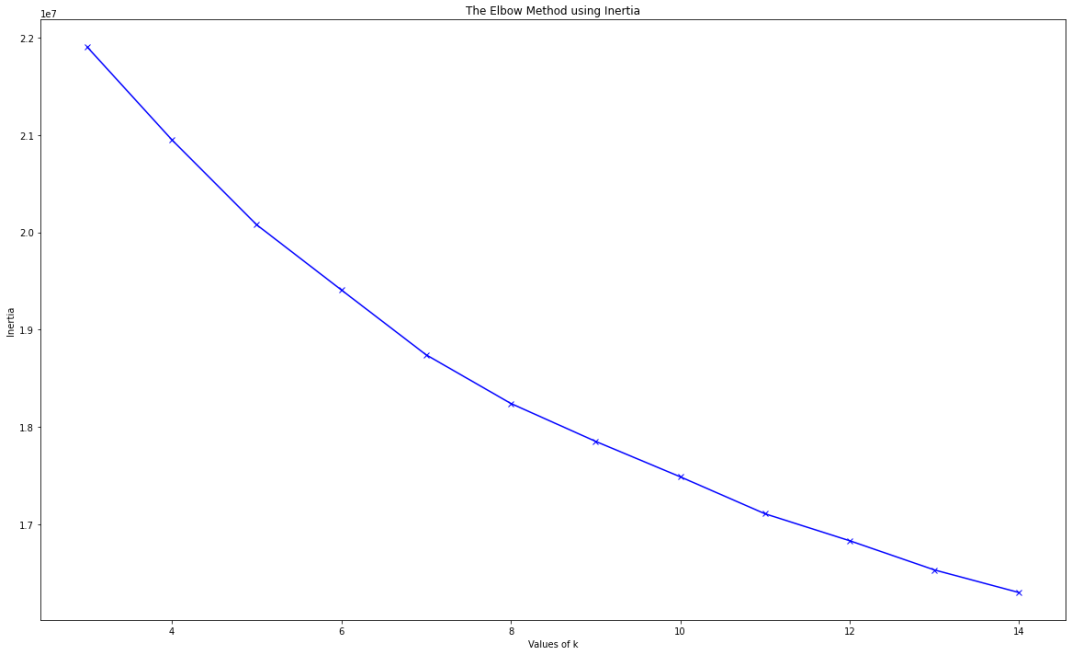


Customer Segmentation

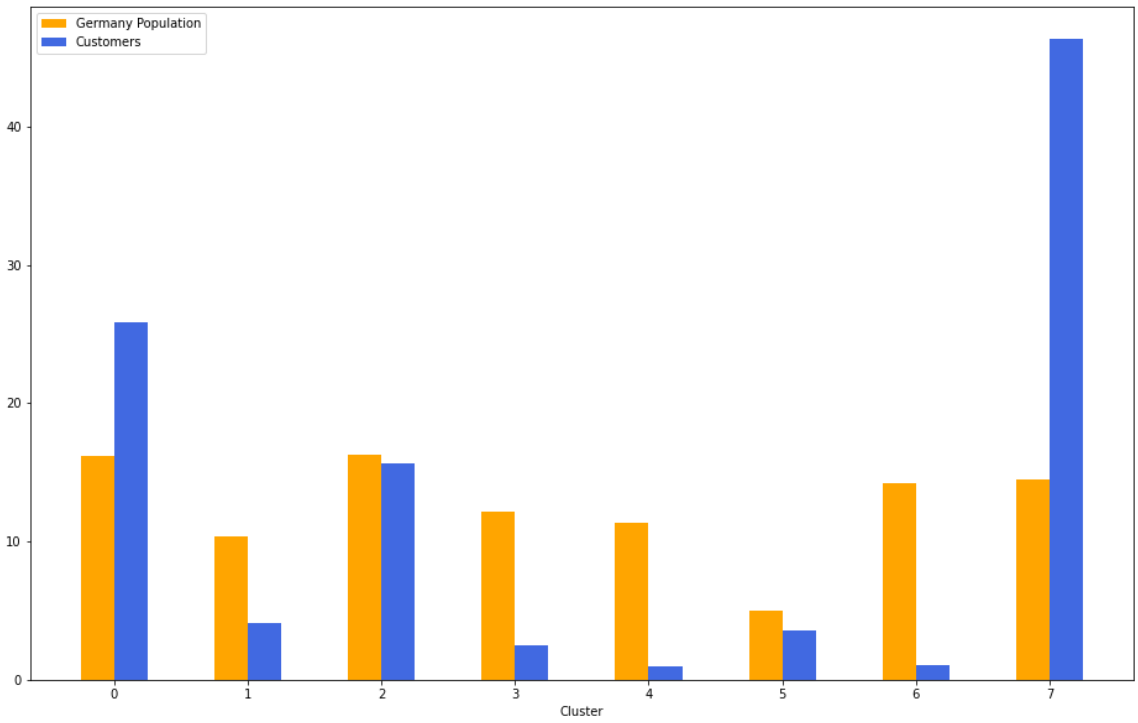
This section describes the relationship between the demographics of the company's existing customers and the general population of Germany. We base our analysis on unsupervised learning methods, namely k-means clustering. We describe parts of the general population that are more likely to be part of the mail-order company's primary customer base.

Clustering

To determine the optimal number of clusters for k-means clustering, we use the inertia *Elbow Method* with several values of k. Recall that the clusters' inertia is defined as the sum of squared Euclidean distances of samples to the closest cluster center. Here are the results using different values of inertia. The following plot shows inertia versus the number of clusters and explains why we chose eight clusters to segment the population.



Next, we projected the CUSTOMERS data set to the principal components of AZDIAS using the model obtained with $k=8$. Notice how the customers populate the different clusters, and clusters 0 and 7 have more customers than the general population.
Clusters 0 and 7 contain about 72% of all customers, and we, therefore, regard them as clusters of interest. We will use these clusters to engineer a feature for the customer acquisition model in part 2.



Demographics of the Centers of Clusters of Interest

To understand the company's customer base, we looked for distinctive features of an average person in clusters 0 and 2. To do this, we used Scikitlearn's **inverse_transform** function to express the centers of these clusters in the original features. Then we identified those features whose values were significantly higher or lower than most of the population using the percentile of the score function. This way, we get characteristics of the center of these clusters that are the most distinctive and help us get a better idea of the company's customer base.

- **Cluster 0**

  An average customer of cluster 0 is likely a rational and culturally minded person who favors environmentally sustainable practices. A typical person in this cluster has a short length of residence, has a high amount of online transactions, does inclined to plan financially.

- **Cluster 7**
  People in this cluster are likely to invest and like to be prepared financially. They have a high length of residence and live household with many adult persons and people with academic titles. They are also not likely to be online and have a high affinity for indicating that they critical minded.

Customer Acquisition

Using supervised learning algorithms, we trained models to predict whether a person would respond positively to a marketing campaign based on the demographic data. First, we cleaned 'Udacity_MAILOUT_052018_TRAIN.csv' using similar cleaning and processing steps for the general population and customer data.

The response column has 42,430 negative responses and only 532 positive. Given how skewed the data set is, we will evaluate the classification models' performance using the AUC-ROC score.

Feature Engineering

To incorporate our customer segmentation results, we predicted each person's clusters in the train and test sets according to the model used in the customer segmentation. We then engineered a feature 'CLUSTER' whose value is one if the individual belongs to the overrepresented clusters (0 and 7) and 0 else.

Feature Selection

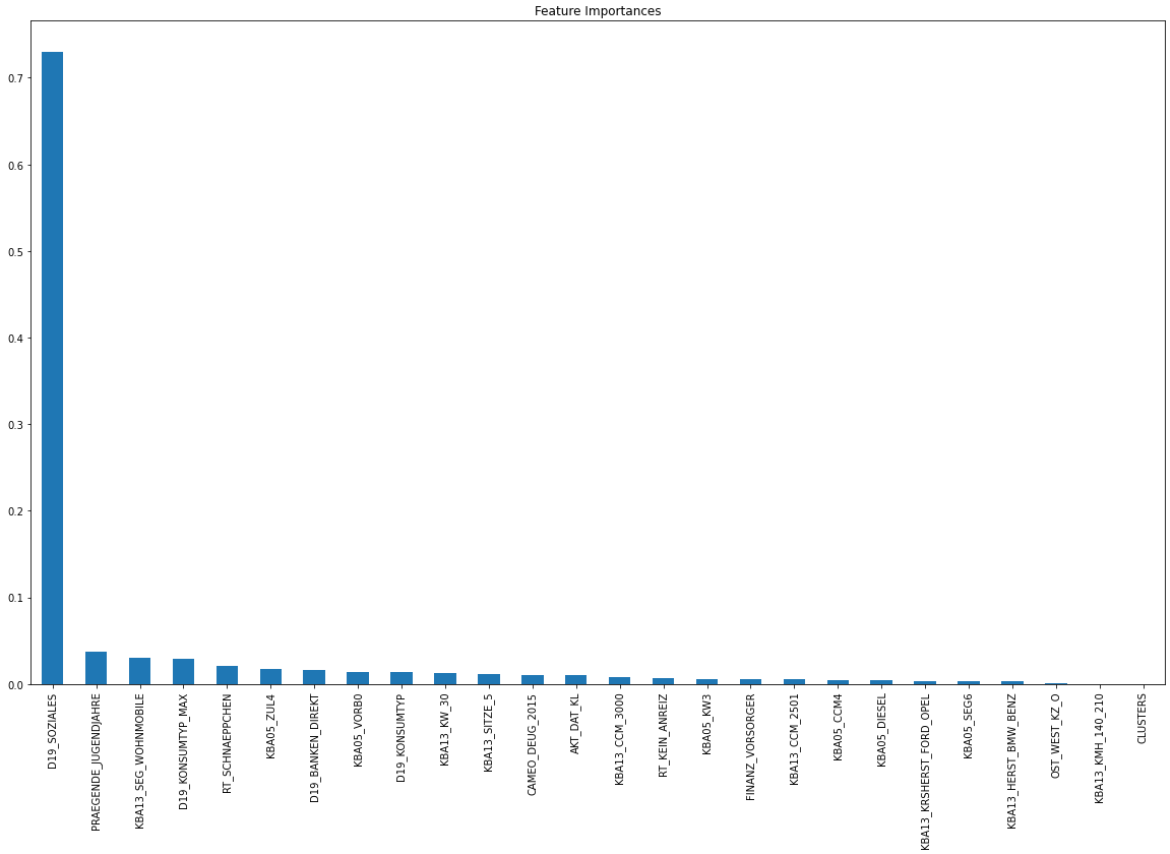
By computing the ANOVA F-value for each feature, we chose the best features based on univariate statistical tests. This preprocessing step to an estimator helped us reduce the number of columns to 26 and improve our classification models' efficiency and performance.

Models

A logistic regression model was trained to obtain a benchmark and fine-tune more sophisticated models. The logistic regression model had an AUC-ROC score of 0.73.

Then we chose a Gradient Boost Classifier (GBC) because of its good performance with imbalanced data. Using the default parameters, the GBC model had an AUC-ROC score of 0.76. We used a cross-validation Grid Search to fine-tune the hyperparameters of the model. The final model had an AUC-ROC score of 0.77 on the test portion of the training set and an AUC-ROC score on the test set which was later submitted to the Kaggle competition.

The most important feature for our model is 'D19_SOZIALES'. Unfortunately, there is no description given in the attribute information files.



Improvements and Future Steps

The distribution of the principal components can be studied more carefully before clustering. Many of the features in the data set were categorical and with somewhat subjective values. If possible, one could merge the demographics data with more objective and factual information and that could potentially lead to a much finer analysis of the customer segments.

Conclusions

This work provides a template to analyze the demographics of a company's customer base and improve customer acquisition through mailout advertising campaigns. The pipeline implemented here consisted of the integration of unsupervised and supervised methods to complement the tasks and give a more colorful story of the data provided.

References

Clustering in Machine Learning
https://developers.google.com/machine-learning/clustering?authuser=1

Scikit-learn
https://scikit-learn.org/stable/