Machine Learning Engineer Nanodegree

Arvato Financial Solutions Customer Segmentation and Prediction Report

Author: Mei Mei

Date: 5/12/2020

OVERVIEW

- **Domain knowledge**
  Arvato Financial Solutions is an international company that helps industry develop innovative solutions for their customers around the world, especially focusing on automation and data analytics.
  Customer segmentation has always been a great tool for company to better tailor their marketing efforts on different group of audience. These differences could be geolocation, demographic and even customer behavior if we can capture them.
  The purpose of marketing segmentation is to generate more profits with less marketing efforts, and CRM (Customer Relationship Management) could also benefits from it. [1]
  This project is specifically working on figuring out the customer base for a mail-order sales company in Germany and then the customer base knowledge gained will be used to increase the return rate of audiences. In this project, by analyzing population attributes and demographic information of two giant datasets, the client can better understand their customer segments and identify future possible customers.

- **Problem statement**
  There are three datasets provided and two problems are raised for machine learning solutions. First one is a clustering problem and the second one is a classification problem.

  Problem 1. Customer Segmentation: By using unsupervised learning methods to analyze attributes of established customers and the general population in Germany to uncover customer's patterns of the mail order company.
  Problem 2. Classification of Customers: Predicting whether a person will be our targeted customer and we should separate the customer from the non-customer as accurate as possible. The model used here will be supervised learning, because we have a third dataset with response that we can make use of and a test dataset to evaluate the model's accuracy.

- **Datasets and inputs**
  The following datasets are adopted in customer segmentation (clustering) problem. And a comparison will be conducted to identify different customer segments.
  1. Udacity_AZDIAS_052018.csv: Demographics data for the general population of Germany; 891 211 persons (rows) x 366 features (columns).
  2. Udacity_CUSTOMERS_052018.csv: Demographics data for customers of a mail-order company; 191 652 persons (rows) x 369 features (columns).
  The following datasets are adopted in customer prediction (classification) problem. And the predicted values will be uploaded to Kaggle for evaluation.

1. Udacity_MAILOUT_052018_TRAIN.csv: Demographics data for individuals who were targets of a marketing campaign; 42 982 persons (rows) x 367 (columns).
2. Udacity_MAILOUT_052018_TEST.csv: Demographics data for individuals who were targets of a marketing campaign; 42 833 persons (rows) x 366 (columns).
There are also two xlsx files for clarification of attributes also coming from Arvato.
(Source: Arvato project workbook.ipynb at Udacity)

- **Benchmark model**
  Logistic regression and Random forest are solid benchmark models for this problem. Upon search I found a fellow classmate achieved 50% accuracy with random forest model. [2] I imagine this could be a good benchmark model for my project. As a popular and widely used classification model, Logistic regression will also be applied as a benchmark model against other more complicated ensemble tree models.

- **Evaluation metrics**
  For the clustering problem, standard errors will be the evaluation metric for finding the best number of clusters with the help of an elbow plot.

  AUROC
  The AUROC is calculated as the area under the ROC curve. A ROC curve shows the trade-off between true positive rate (TPR) and false positive rate (FPR) across different decision thresholds. In the second problem here, it is an imbalanced dataset, AUROC (Area Under Receiver Operating Characteristic) is selected as an evaluation metric for the classification problem. The rationale behind this is because AUROC considers both true negative and true positive in the group. with 42430 negative and only 532 positives. As for evaluation metric for imbalanced data prediction, we need to carefully select because accuracy won't get us what we want, e.g. with labeling all the response as negative, the model will get an accuracy of (42430-532)/42430 is almost 100%. Obviously, we are not happy with a model like that. We want a model to better divide positive from negative, thus AUROC is highly recommended for both problems here.

- **Project design**

  Data exploration and preprocessing: initial analysis on the data showed a significant amount of missing data, further analysis on missing rows and columns will be done. Type of attribute (Numeric, categorical, ordinal etc.) should be carefully handled as well.

  Customer Segmentation: Data Transformation, PCA, K-Means Clustering, AZDIAS/CUSTOMER Data Comparison. Customer Prediction: Data Transformation, Classifier evaluation, Classifier Parametrization.

ANALYSIS

- **Solution statement**
  First all data needs to be cleaned, missing values and columns should be handled in a reasonable way.
  Dimensionality reduction technique PCA will be used for Customer segmentation to attain the minimum number of dimensions that explains as much variation as possible. A clustering algorithm like K-Means will be applied for segmenting the customers into different clusters.
  The second part will be doing supervised machine learning on the training and testing data for finding potential customers. Because the training data is clearly an unbalanced data (Only 532 observations are labeled as 1), in order to not lose any more positive data, I will not delete any more row data like I do in the first part of this project. Different ensemble tree models including random forest, gradient boost tree and Adaboost tree will be trained for this problem and tested on validation set.

- **Data Cleaning**
  The table shows the shape of both datasets – The general population and customers, and I will refer them as Azdias and Customers respectively for the remaining of the report.

Number of population rows	891221
Number of population columns	366
Number of customers rows	191652
Number of customers columns	369

  1. Convert invalid values 'X' and 'XX' to Null for columns '*CAMEO_DEUG_2015'*, '*CAMEO_INTL_2015'* and '*CAMEO_DEU_2015'*
  2. In DIAS Attributes dataset I found out that missing data adopts different values: -1, 0, 1, 9 etc. With function missing_value() I was able to convert all missing data as Null value.
  3. Delete columns with too much missing data. Column that has more than 30 percent of missing data will be removed from further analysis to avoid introducing too much imputation. In this step, 17 features were removed.

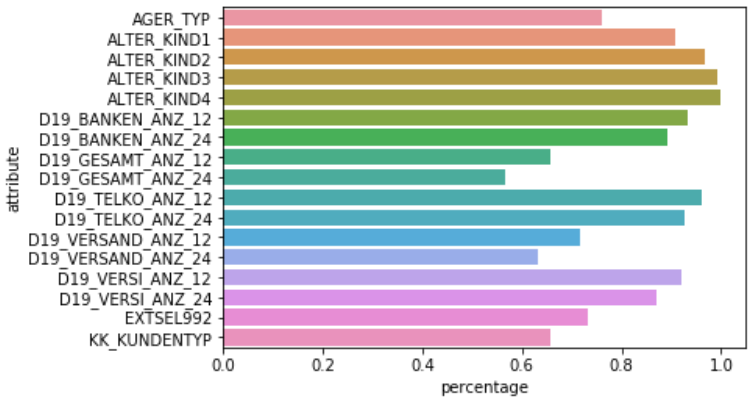


Figure 1. (>30 missing values) Attributes in Azdias data

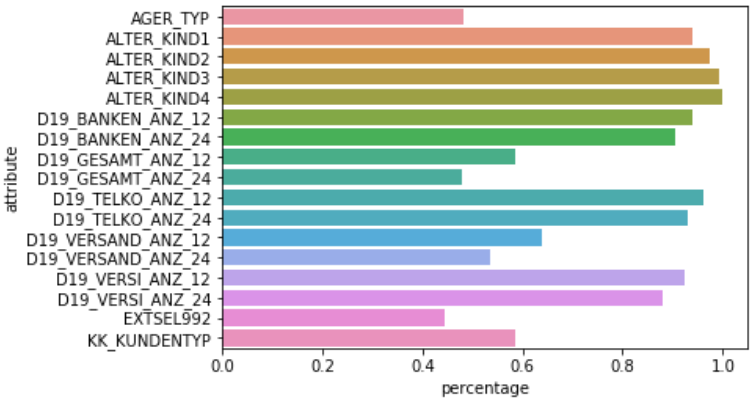


Figure 2. (>30 missing values) Attributes in Customers data

4. Some variables have a significant amount of unique values, cleaning and grouping them will help with clustering later. Affected columns here are: *'LP_LEBENSPHASE_FEIN'* (age level) and '*LP_LEBENSPHASE_GROB'* (income level). They used to have 40 categories, but after regrouping, I was able to label all the values with only four variations.
5. Date data cleaning for column '*EINGEFUEGT_AM'*, same as last step, I grouped them by years in order to reduce categories.
6. Re-label values from {W, O} into {0, 1} for *'OST_WEST_KZ'*
7. Drop the gap attributes of Azdias and customers dataset which are *'PRODUCT_GROUP' 'CUSTOMER_GROUP' and 'ONLINE_PURCHASE'.*

New shape of datasets

Number of population rows	731331
Number of population columns	347
Number of customers rows	135144
Number of customers columns	347

- **Imputation and Standardization**

Missing values are replaced by most frequent values. PCA is affected by scale so you need to scale the features in the data before applying PCA. The data is transformed onto unit scale (mean = 0 and variance = 1) which is a requirement for the optimal performance of many machine learning algorithms. StandardScaler helps standardize the dataset's features. It is fit on the training set and transform on the training and test set.

- **Dimension reduction**
  We have 347 features in total in general population and customers. However not all of them have meaningful variation, i.e. Collinearity. Also, tiny variation within the feature could happen, even some features will remain the same for all the people. A systematic approach should be addressed to perform dimensionality reduction. I leveraged Principle Component Analysis (PCA) to decide which feature to keep, which feature to drop.

MODELS

- **Dimension reduction**
  The gist of using PCA is to apply a linear transformation on the data and get the ratio of explained variance. By setting a threshold, for my case here, I selected 90%, that means 90% variance is explained by all the selected features. And I was able to get 168 dimensions as my final components.

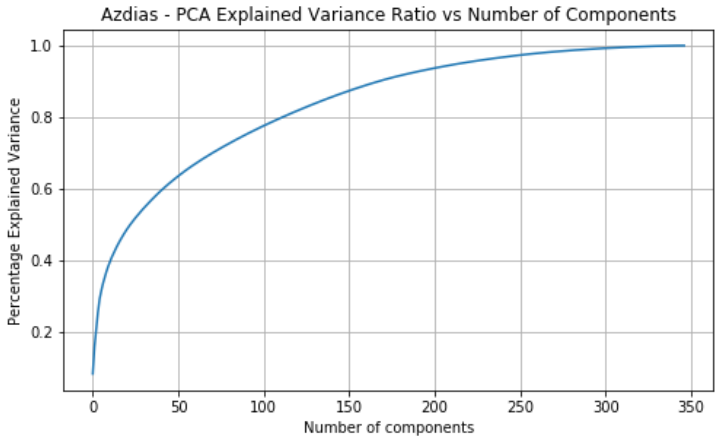


Figure 3. Azdias-PCA explained variance ratio vs number of components

- **Components analysis**

  Component 1:

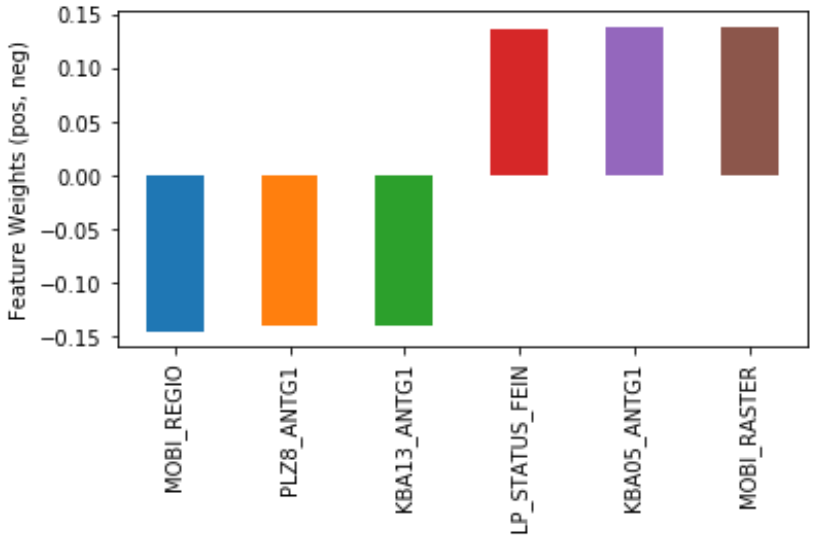


Figure 4. Most important features for component 1

Positive impact: High Income
Negative Impact: Frequent Movers; Sharing house with other families

Component 2

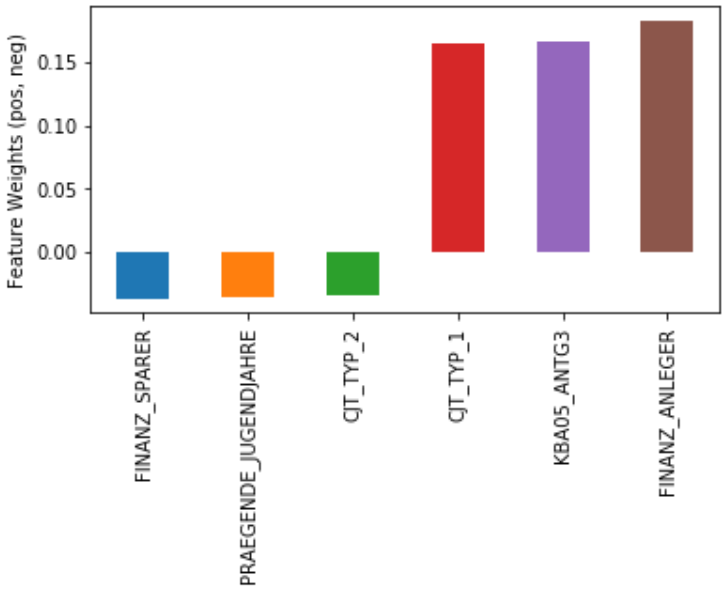


Figure 5. Most important features for component 2

Positive Impact: Low Investing party
Negative Impact: Money Saver; Older Generation

Component 3

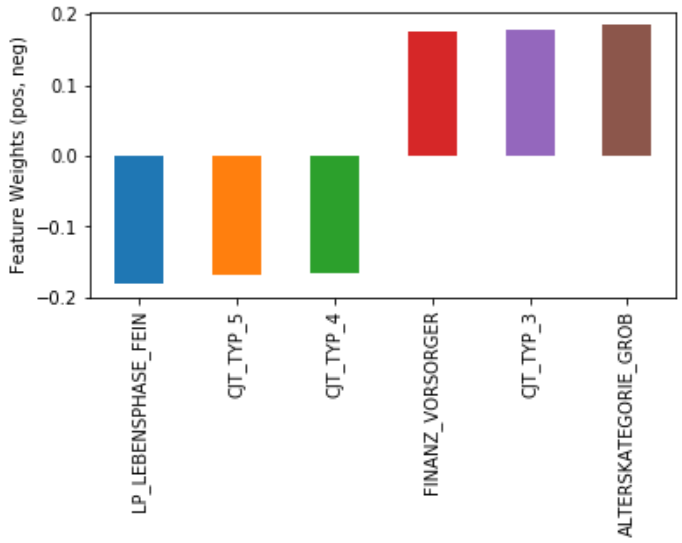


Figure 6. Most important features for component 3

Positive Impact: Financial Prepared
Negative Impact: CJT_TYP_5 and CJT_TYP_4 (Cannot find explanation)

- **Clustering**
  Elbow function is adopted to decide the best number of clusters.

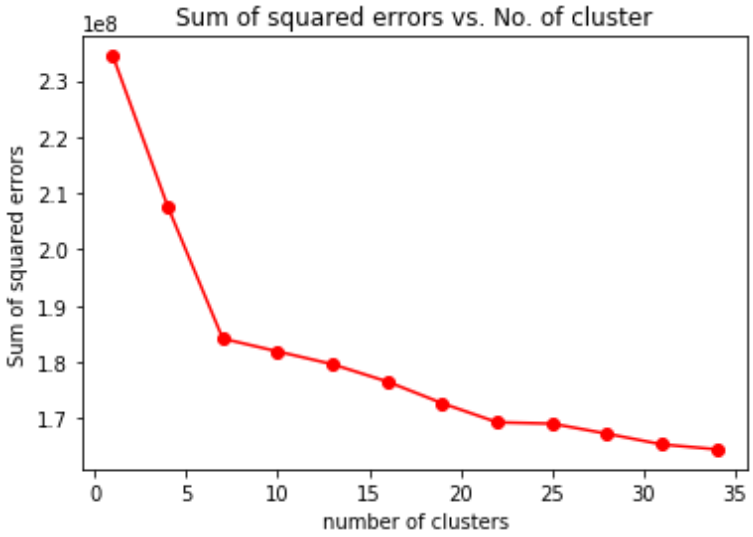


Figure 7. Sum of squared errors vs. No. of clusters

Based on the elbow graph above, I chose 16 as a reasonable number of clusters.

- **Classification**
  Conduct the same steps of data cleaning on mailout_train dataset. The Mailout data is split into two equal parts, each containing about 43,000 rows of records.
  Four models are trained on the training dataset. There are Logistic Regression, Random Forest Classifier, AdaBoost Classifier and Gradient Boosting Classifier.

After calibration of the best performing model on the training data, prediction is done on the test data, and result is uploaded to Kaggle.

CONCLUSION

- **Clustering**

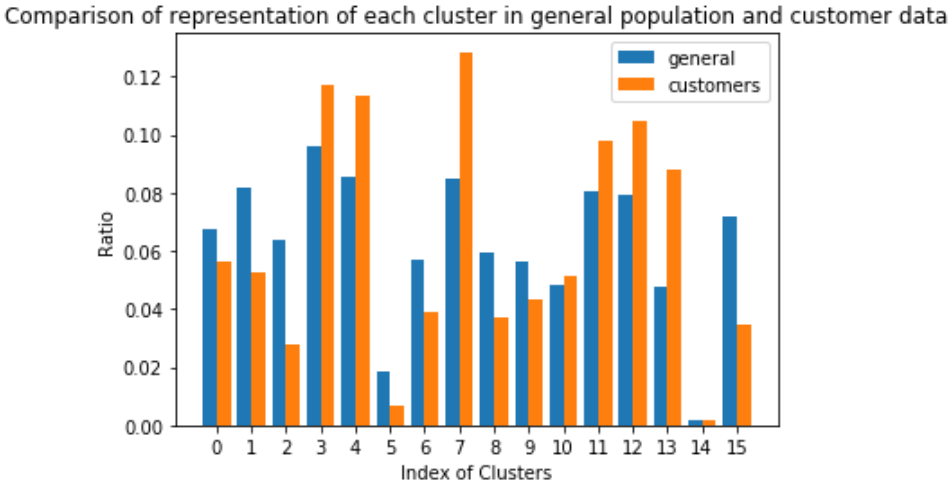


Figure 8. Comparison of representation of each cluster in general population and customers data

Based on the comparison graph above, we can tell that cluster 7, 13 are overrepresented in customers dataset, and cluster 2, 8, 15 are underrepresented in customer dataset. I will investigate each of these clusters respectively.

Over Representation Cluster with most two important components

Cluster 7		**Cluster 13**	
Weights	Component	Weights	Component
0.959620	1	1.826480	5
0.948558	3	1.611521	7

Based on the components above, we can get some insights about characteristics of existing customers:

1.High income with less frequent home moving.
2.High transaction of mail-order over the past 12 months.
3.High transaction activity TOTAL POOL in the last 12 months.
4.People who is less critical minded, less event-oriented, more dominant minded and more with a fightful attitude.

5.Family shopper, demanding shopper.

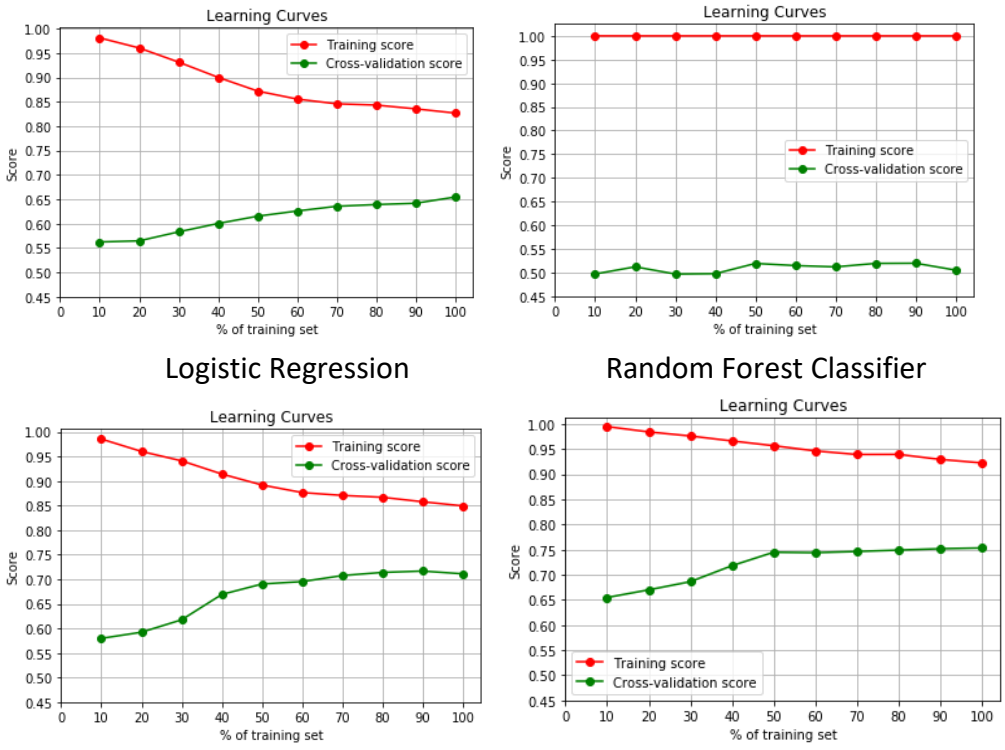
Under Representation Cluster with most two important components

Cluster 2		Cluster 8		Cluster 15	
Weights	Component	Weights	Component	Weights	Component
5.596022	3	4.847072	0	4.809255	0
5.079054	0	2.384620	2	2.444621	2

Based on the components above, we can get some insights about population who are less likely to be our customers:

1.People who cares about environmental sustainability in the youth
2.People who are more a money saver
3.People who has higher share of 1-2 family homes
4.People who is less likely an investor
5.People at younger age

- **Classification**



Logistic Regression Random Forest Classifier

AdaBoost Classifier Gradient Boosting Classifier

Figure 8. Learning curves comparison of four models

Model	Roc_auc train score	Roc_auc test score
Logistic Regression	0.83	0.65
Random Forest Classifier	1.0	0.51
AdaBoost Classifier	0.85	0.71
Gradient Boosting Classifier	0.92	0.75

- **Model Comparison**
  Gradient Boosting Classifier has the best performance of 0.75 on test data, and good performance on training data. It is my target model for Kaggle submission.
  Logistic Regression has a solid performance and since two curves are not converging yet, I believe with more training and testing data, we can achieve better results from this model.
  Random forest shows overfitting pattern with high AUROC on training data and AUROC (50%) on testing data.
  AdaBoost Classifier has an overall good performance.

With Grid search, I can find the best hyper parameters for Gradient boost classifier, and the final Roc_auc score on training data is 0.8908.

- **Kaggle Competition**
  To predict the result of a dataset, first I cleaned mail_test data same as before, then I used the saved best Gradient boost classifier model on mail_test data.
  I got score of 0.79140 at Kaggle, the best submitted result on Kaggle is 0.81063.

Reference

[1] Kim, S. Y., Jung, T. S., Suh, E. H., & Hwang, H. S. (2006). Customer segmentation and strategy development based on customer lifetime value: A case study. *Expert systems with applications*, *31*(1), 101-107.
[2] https://towardsdatascience.com/customer-segmentation-report-for-arvato-financial-solutions-b08a01ac7bc0