# Machine Learning Engineer Nanodegree

Customer Segmentation Report - Arvato Financial Solutions

Hongwei Niu

October 26, 2020

# 1 Definition

## 1.1 Project overview

In this project, we will work with a dataset from a mail-order sales company in Germany provided by Udacity and Arvato [1]. It includes general population dataset, customer segment data set, dataset of mailout campaign with response and test dataset that needs to make predictions. The goal is to characterize customers segment of population and to build a model to predict if a customer will response to the campaign.

## 1.2 Problem Statement

The ultimate goal of this project is to answer this acquisition question. Given the demographic data, how a mail-order company could acquire new customer effectively? There are two main parts to solution this problem.

First, an unsupervised machine learning algorithm will be applied to general population data and customer data for user segmentation. This will help us to identify the demographic feature of the users that mostly likely to become customers.

More specifically, we will use K-MEAN algorithm to perform customers segmentation of general population and then to describe the cluster of population who most likely convert. As we all know that one of the limitation

of K-mean is high dimensional data. To solve this problem, we will use PCA to reduce data dimensionality first.

Second, a supervised learning algorithm will be used to predict if a person could turn to a customer based on the demographic data. This is a typical binary classification problem. We will choose several binary classification algorithms, then compare their performances, finally choose the best one.

### 1.3 Metrics

1. Unsupervised Leaning algorithms.

As mentioned above, PCA and K-Mean will be used for user segmentation. PCA is a common technique for dimentionality reduction. The idea is that using the minimal number of dimensions to explain the variance as much as possible. Therefore, the explained variance ratio will be used to help choose the minimal number of dimensions. Next, in order to choose the best number of clusters for K-Mean algorithm, the distance of Elbow method was used to identify an ideal number of clusters for k-means clustering on the PCA-transformed data, and the average of sum of squared errors (SSE) will be used for this purpose.

2. Supervised Leaning algorithms.

One of the most common evaluation metrics for binary classification problem is Area Under ROC Curve [2]. It presents the ability of a model to discriminate between positive and negative classes. An area of 1.0 represents a model could predict positive classes perfectly. And an area of 0.5 represents that the prediction of a model is as good as random. It will be used as the evaluation metric for supervised Leaning algorithms

## 2 Analysis

### 2.1 Data Exploration and Preprocessing

1. Remove columns with over 50% of NaN

After loading the datasets, we notice that the dataset Azdias contains lots of NaNs. We shall remove the columns that with over 50% NaNs since they do not have enough meaningful information.

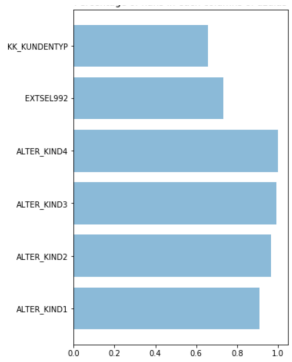


Figure 1: Columns with over 50% missing values

2. Replace unknow with NaN

After exploring the two atrributes information datasets, we notice that there are some values in attributes which are marked as unknow. We replaced them with NaN.

3. Fix columns having mixed data types

There are two columns with mixed d types: CAMEO_DEUG_2015 and CAMEO_INTL_2015. This is because there are two records X and XX which are not numerical. We shall set them to NaN.

4. Encoding Data

Since most machine learning models only accept numerical variables, we shall convert categorical variables to numbers so that the model is able to understand and extract valuable information. The encoded columns are D19_LETZTER_KAUF_BRANCHE, EINGEFUEGT_AM, OST_WEST_KZ, PRODUCT_GROUP, CUSTOMER_GROUP.

5. Missing values

We filled these nulls with mean.

6. Data normalization

A standard scaler is used to make all the feature into the same range. This is a procedure to avoid feature dominance when using PCA.

# 3 Algorithms and Techniques and methodologies

## 3.1 Customer segmentation

Recall that the aim of this part is to divide the general population and customers into segments in order to compare these two groups to describe the future customers. There are 360 features in each dataset after precossing. Considering the computational complexity, it is import to find which features carries the most information.

## 3.2 PCA

PCA is a method combines highly correlated variables together to form a smaller number of an artificial set of variables which is called "principal components" that account for most variance in the data. Applying PCA to the processed dataset, the plot in Figure 2 shows the cumulative variance explained by principle components.

It shows that 90% variance explained by 160 components. So we use 160 features for the later k-Mean clustering. Next, we shall see what are the most important features that the top 2 principle components correspond.

The component in Figure 3 is associated with people's age, social status and wealth. And the component in Figure 4 is associated with people's online transactions. More components are explained in the jupter notebook.

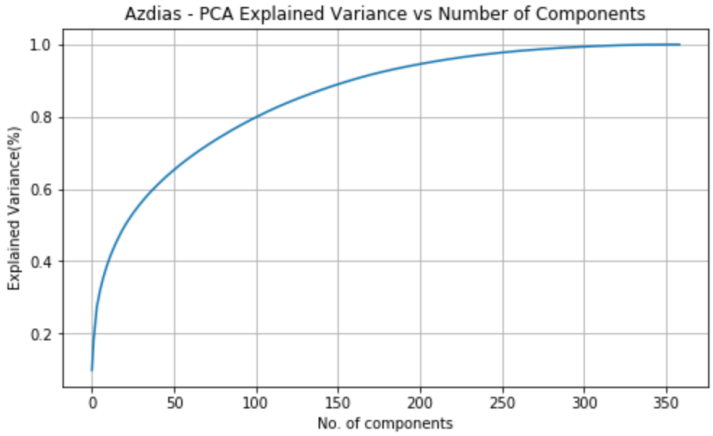


Figure 2: Variance Explained by Components

| | Feature | Description | FeatureWeight |
|---|---|---|---|
| **2** | ALTERSKATEGORIE_GROB | age through prename analysis | 0.271222 |
| **1** | LP_STATUS_FEIN | social status fine | 0.166222 |
| **0** | LP_STATUS_GROB | social status rough | 0.161264 |
| **5** | PLZ8_ANTG4 | number of >10 family houses in the PLZ8 | -0.136024 |
| **4** | KBA13_ANTG4 | No description given | -0.139688 |
| **3** | KBA13_BAUMAX | No description given | -0.142606 |

Figure 3: PCA - Component 0

| | Feature | Description | FeatureWeight |
|---|---|---|---|
| **2** | ALTERSKATEGORIE_GROB | age through prename analysis | 0.167894 |
| **1** | D19_GESAMT_ONLINE_QUOTE_12 | amount of online transactions within all trans... | 0.149928 |
| **0** | EWDICHTE | density of inhabitants per square kilometer | 0.147904 |
| **5** | D19_GESAMT_ONLINE_DATUM | actuality of the last transaction with the com... | -0.168027 |
| **4** | D19_GESAMT_DATUM | actuality of the last transaction with the com... | -0.176263 |
| **3** | D19_KONSUMTYP_MAX | No description given | -0.180266 |

Figure 4: PCA - Component 1

## 3.3 K-Means Clustering

Next, we use unsupervised method K-Means to divide the population and customers into different segments. The hyper-parameter for K-Means is the number of clusters. The idea for choosing the number of clusters is to have the minimal distance of the data points within a cluster. The average of sum of squared errors (SSE) within-cluster distances is chosen for this purpose. we use Elbow to see SSE vs the number of clusters. The plot Figure 5 shows that the score decreased rapidly for the first 8 clusters, then continue decreasing slowly. Therefore, we choose 8 to be the number of clusters.

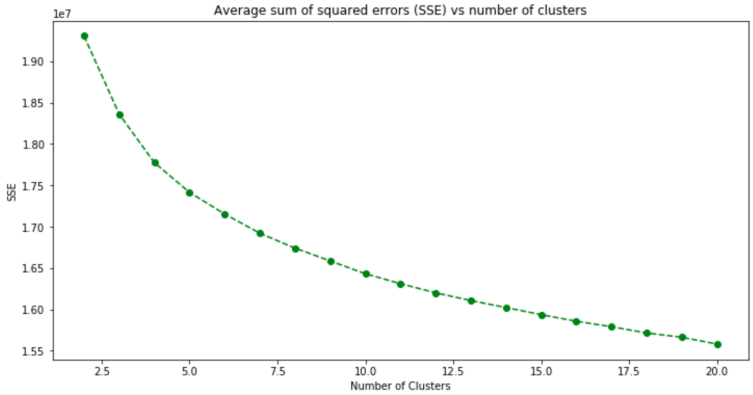


Figure 5: SSE vs No. cluster

## 3.4 Segments

After applying K-Means with 8 clusters, the result is demonstrated in Figure 6, which are the distributions of general population and customers in each of the 8 clusters. Clearly the cluster of distribution of the general population is nearly uniform. And the customers seems from cluster 3 and cluster 7 since the proportion of customers in these two clusters are higher than others .

Moreover, the proportion ratio of general population segments and customers' segments confirms this conclusion, see Figure 7.

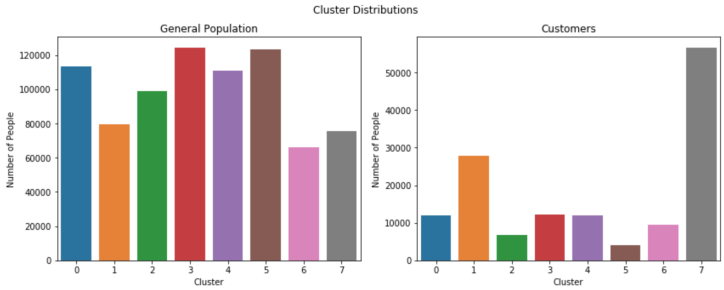


Figure 6: Cluster proportion

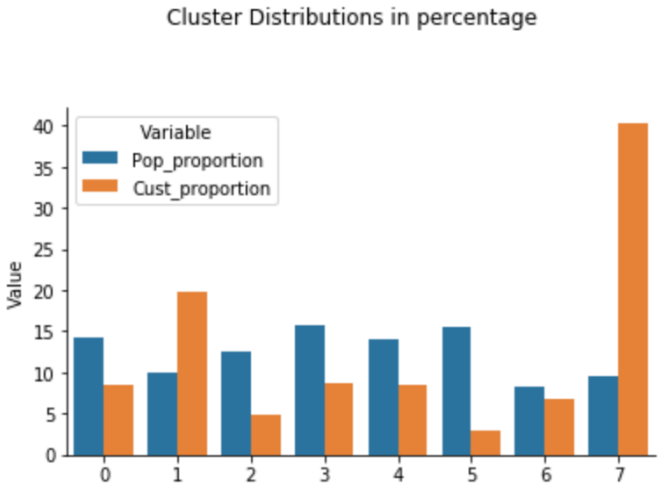


Figure 7: Cluster ratio

## 3.5 Cluster analysis

By conducting cluster analysis, we are able to have a deeper look at each cluster. The result of Cluster 7 is shown in Figure 9, its main component is Component 0 and main features are people's age and wealth.

| Component | ComponentWeight | Feature | Description | FeatureWeight |
|---|---|---|---|---|
| 0 | 0.94932 | ALTERSKATEGORIE_GROB | age through prename analysis | 0.271222 |
| 0 | 0.94932 | LP_STATUS_FEIN | social status fine | 0.166222 |
| 0 | 0.94932 | LP_STATUS_GROB | social status rough | 0.161264 |
| 0 | 0.94932 | PLZ8_ANTG4 | number of >10 family houses in the PLZ8 | -0.136024 |
| 0 | 0.94932 | KBA13_ANTG4 | No description given | -0.139688 |
| 0 | 0.94932 | KBA13_BAUMAX | No description given | -0.142606 |
| 0 | 0.94932 | ALTERSKATEGORIE_GROB | age through prename analysis | 0.167894 |
| 0 | 0.94932 | D19_GESAMT_ONLINE_QUOTE_12 | amount of online transactions within all trans... | 0.149928 |
| 0 | 0.94932 | EWDICHTE | density of inhabitants per square kilometer | 0.147904 |
| 0 | 0.94932 | D19_GESAMT_ONLINE_DATUM | actuality of the last transaction with the com... | -0.168027 |
| 0 | 0.94932 | D19_GESAMT_DATUM | actuality of the last transaction with the com... | -0.176263 |
| 0 | 0.94932 | D19_KONSUMTYP_MAX | No description given | -0.180266 |

Figure 8: Cluster 7

## 3.6 Classification

The final problem is to predict who would become a customer based on the demographic data. The dataset of mailout campaign with response is our training dataset. First of all, we need to prepare the training data by applying to the first 5 steps mentioned in Section 2.1 Data Exploration and Preprocessing.

Since this is a typical classification problem. We shall use the following algorithms and choose the best one to do prediction.

1. Logistic Regression
2. Decision Tree Classifier
3. Random Forest Classifier
4. Gradient Boosting Classifier
5. AdaBoost Classifier

And Gradient Boosting Classifier has the best AUCROC_score shown in Figure 9.

| | Model | AUCROC_score |
|---|---|---|
| 0 | LogisticRegression | 0.67614 |
| 1 | DecisionTreeClassifier | 0.486943 |
| 2 | RandomForestClassifier | 0.501053 |
| 3 | GradientBoostingClassifier | 0.77585 |
| 4 | AdaBoostClassifier | 0.754274 |

Figure 9: ROC score

### 3.7 Benchmark

The benchmark chosen for this prediction task is the highest score in the Kaggle competition, which is 0.84739. However, the best performance we have is 0.77585.

### 3.8 Refinement

Recall that the training dataset previousely used is neither normalized nor dimensionality reduced by PCA. We shall normalize the training dataset and reduce dimensionality by PCA to see if the classifiers' performance could be improved.

As shown in Figure 10, this method could not improve models performance. We shall use the first Gradient Boosting Classifier to prediction the mail-out test dataset without scaling or reducing dimensions.

## 4 Result

Applying Gradient Boosting Classifier for prediction on the mail-out test dataset, we submit the result to Kaggle. The final score is 0.79379.

| | Model | AUCROC_score | AUCROC_score_scaled | AUCROC_score_pca |
|---|---|---|---|---|
| 0 | LogisticRegression | 0.67614 | 0.659056 | 0.67354 |
| 1 | DecisionTreeClassifier | 0.486943 | 0.481687 | 0.490414 |
| 2 | RandomForestClassifier | 0.501053 | 0.503311 | 0.491401 |
| 3 | GradientBoostingClassifier | 0.77585 | 0.773839 | 0.566981 |
| 4 | AdaBoostClassifier | 0.754274 | 0.749613 | 0.550982 |

Figure 10: ROC score with scaled and dimention reduced training set

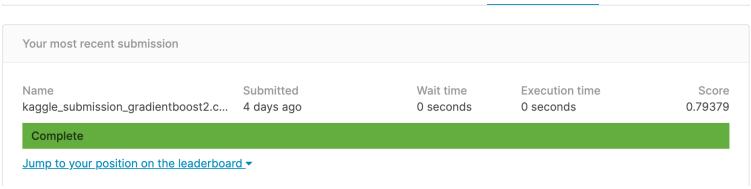


Figure 11: Kaggle competition result

# 5 Improvements

The score that we achieved is not ideal. There are approaches that might improve the Kaggle competition score.

1. Understand more features in the dataset and select most relevant ones.
2. Observe that the training dataset is imbalanced. Up-sampling, down-sampling methods could be applied.

# 6 References

1. Arvoto, https://www.bertelsmann.com/divisions/arvato/st1

2. A New Method for Dimensionality Reduction using K-Means Clustering Algorithm for High Dimensional Data Sets Clustering Algorithm for High Dimensional Data Set

3. Metrics To Evaluate Machine Learning Algorithms in Python