Udacity Machine Learning Nanodegree

Customer segmentation and Mailout's response prediction report

Korntewin Boonchuay
10 May 2021

Definition

Domain Background

Arvato is a data-driven organization and an internationally services company who provide various solutions to the business customers, including Supply Chain Management, financial services and IT services which continuously developed using the power of data and analytics [1].

Interestingly, there are around 200,000 existing Arvato's customers and 360 demographic features for each individual. With this much data, it is almost impossible to manually explore and identify customer's insight. Thus, data mining and machine learning techniques will be an integral part in finding customer's insight. Additionally, with the mindset of data-driven organization, the results from data analytics will lead to business action and making impact to Arvato business.

Market segmentation technique is suitable for the current Arvato business both in terms of increasing existing customers satisfaction [2] and effectively approaching potential customers. Thus, in this project, various machine learning techniques will be implemented to cluster and identify the potential customers. The underlying latent and insight from market segmentation will also be used to predict mailout responses for each individual and improve the mailout strategies for the Arvato marketing team.

Problem Statement

To cluster the population and identify the potential customers, the data of the general population is needed. In this project, AZ DIAS information database [3] is provided as a general population data with the strict terms of use.

The project will be divided into 2 main tasks as follows.

Customer segmentation

The unsupervised machine learning technique will be implemented to cluster the population using general population data/Arvato's existing customers data and identify the demographic pattern of the potential customers.

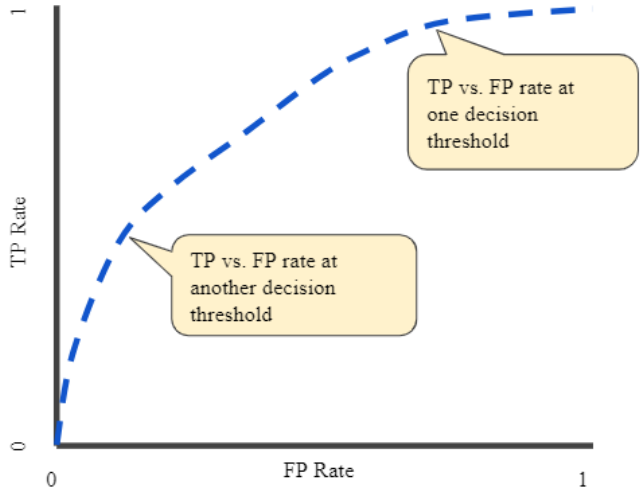
Mailout's response prediction

The underlying latent and information from the previous customer segmentation task together with existing mailout data will be used to effectively classify the population, who are likely to respond to the Arvato's mailout, using supervised machine learning.

Evaluation Metrics

Mailout response prediction task

Since the label for the binary classification tasks is highly imbalanced, the area under the curve of the receiver operating characteristic curve (AUC ROC) will be used to evaluate the model. The AUC ROC can be calculated from the area of the True Positive Rate and False Positive Rate at various decision thresholds as shown below.



Receiver operating characteristic curve [4]

AUC ROD is ranging from 0 to 1. One hundred percent accuracy model has an AUC ROC of 1.0 and zero percent accuracy model has an AUC ROC of 0.0. However, the practical benchmark AUC ROC is 0.5.

This metric is suitable for the task with an imbalance label. For instance, on a task with 99% positive label, the accuracy metric will be easily biased to 99% with majority guessing. But the majority guessing or random guessing will only result in **0.5 of AUC ROC**.

Analysis

Datasets and Inputs

There are 4 given files containing demographic data for Arvato's existing customer and general population in Germany provided by Arvato Bertelmann as follow:

1. Udacity_CUSTOMERS_052018.csv
    a. This file contains demographic data for Arvato's existing customers of 191,652 individuals with 369 features.
2. Udacity_AZDIAS_052018.csv
    a. This file contains demographic data for the general population in Germany of 891,221 individuals with 366 features.
3. Udacity_MAILOUT_052018_TEST.csv
    a. This file contains demographic data for the population who were targeted by a marketing campaign of 42,833 individuals with 366 features.
4. Udacity_MAILOUT_052018_TRAIN.csv
    a. This file contains demographic data for the population who were targeted by a marketing campaign of 42,982 individuals with 367 features.

Furthermore, there are 2 data dictionary or metadata files describing the information, all possible values and encoding null value of each features as follow:

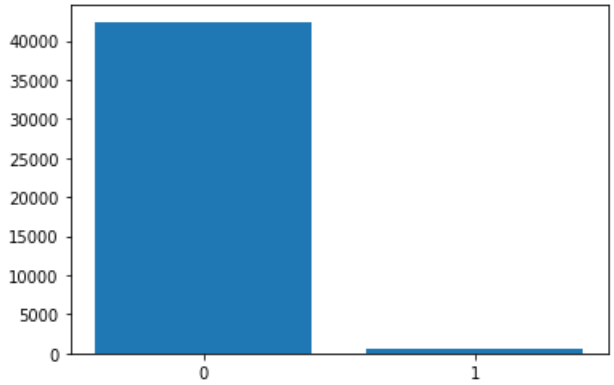
1. DIAS Attributes - Values 2017.xlsx
    a. This file contains all possible values and how null value is encoded for each feature.
2. DIAS Information Levels - Attributes 2017.xlsx
    a. This file contains the description of each feature.

These individuals data will be explored, feature engineer and used as an input for customer segmentation and mailout response prediction.

Exploratory Data Analysis

Label imbalance observing

First, the balance of the label in mailout datasets should be observed. Since it is the most important thing in choosing metrics to evaluate the classification model. The number of label for "0" (no response) and "1" (does response) can be illustrated as below:



There is only 1.2% of the label “1” which indicates the heavy imbalance for classification models. Thus, the metrics for evaluating should be precision, recall, F1 score or Area Under Curve of ROC.

Level of measurement observing

All of the features will be manually categorized to be one of the following levels:

1. Nominal: Categorical data which its semantic cannot be sorted, such as the type of animal.
2. Ordinal: Categorical data which its semantic can be sorted, such as student grade.
3. Interval and ratio: Numerical data
4. Temporal data: The data that represent the state in specific time.

Fortunately, almost all of the features can be categorized by data catalog file “DIAS Attributes - Values 2017.xlsx” and “DIAS Information Levels - Attributes 2017.xlsx”. Almost 90% of the features are categorized as ordinal. The meta data of each feature is saved in json format.

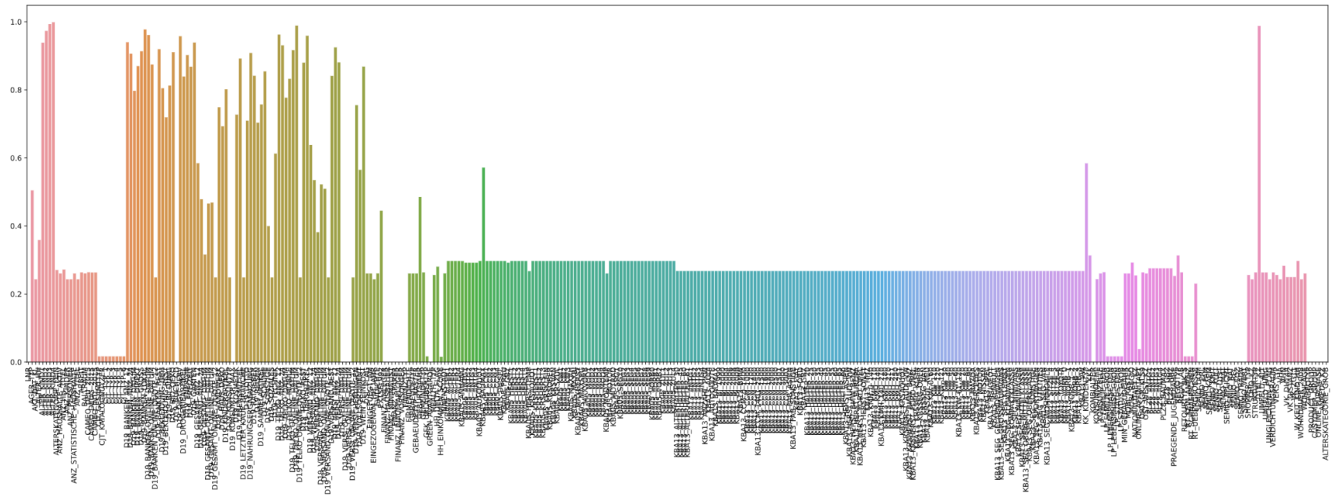
Feature missing percentage observing

The feature with too much missing value will degrade the model performance for the following reasons:

- The feature may be quasi constant. For instance, if some feature has 90% missing value and it is imputed with the median value, then there will be 90% of the same value (median) in such feature.
- The remaining value in the feature may not be suitable to represent the missing value using statistical calculation.

Thus, the feature with too many missing values will be filtered out. In this project, the missing percentage threshold is set to be 30%, meaning that the feature with missing percentage more than 30% will be filtered out. Note that the different features have different encoded missing values.

The missing percentage of each feature can be illustrated below:



Some features, such as ALTER_KIND1 - ALTER_KIND4 have a missing value of more than 90%. While the majority of the missing percentage is around 20%.

Record missing percentage observing

Similar to feature missing value, the incomplete record or incomplete individual data can also degrade the model performance. The incomplete record can mislead the model because many features in that record have to be imputed and cause overfitting.

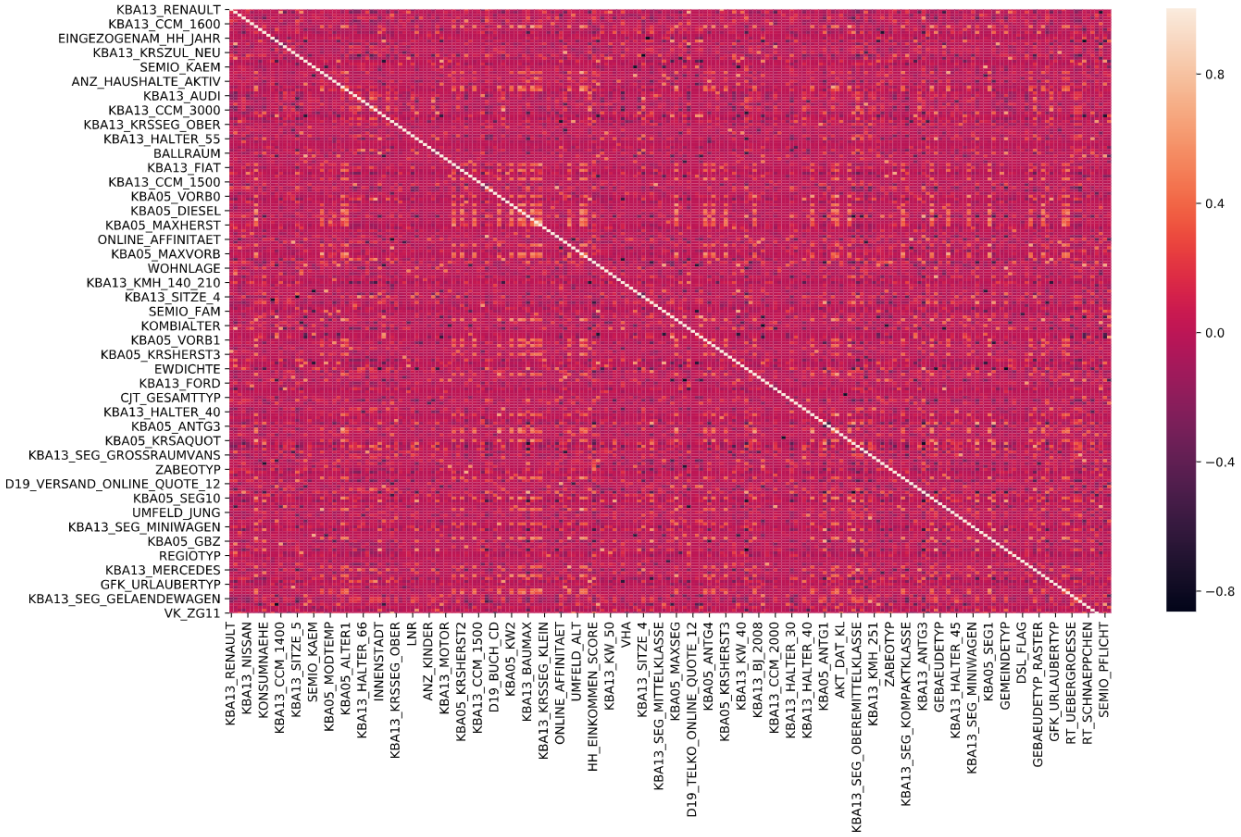
By setting threshold of 30%, the percentage of the record that exceed the threshold can be shown below:

Udacity_AZDIAS_052018.csv	15.69%
Udacity_CUSTOMERS_052018.csv	28.51%
Udacity_MAILOUT_052018_TRAIN.csv	20.01%

Feature correlation observing

The feature that is correlated with another feature should not be included together in the model. The redundancy of many features will result in instability of the estimated parameters and makes it difficult to evaluate the effect of the feature to the output of the model [5].

In this project, correlation threshold of 0.7 is set and the feature correlation matrix after filtering using this criteria can be illustrated below:



Cardinality observing

The nominal level features that contain too many distinct values should be filtered out. Because each distinct value will be spanned out as one feature using one hot encoder. The input feature of the model will be too sparse and cause the model to overfit.

By setting threshold of 15 distinct values, the feature that exceed this threshold are as follow:

Udacity_AZDIAS_052018.csv	['CAMEO_DEU_2015', 'CAMEO_DEUG_2015', 'LP_LEBENSPHASE_FEIN', 'PRAEGENDE_JUGENDJAHRE']
Udacity_CUSTOMERS_052018.csv	['CAMEO_DEU_2015', 'CAMEO_DEUG_2015', 'LP_LEBENSPHASE_FEIN', 'PRAEGENDE_JUGENDJAHRE']

Quasi-constant observing

The features that are almost constant for all of the record will be filtered out to reduce overfitting, since such features will not give much information. By setting the threshold to 90%, there are around 20 features that exceed this threshold.

Algorithms and Machine Learning Techniques

The implemented data cleaning and machine learning techniques to achieve project goals can be described as follow:

Data cleaning

The data cleaning pipeline is consisted of:

1. Column selecting transformer to select only the selected feature
2. Encode unknown value transformer to convert the missing value to NaN
3. Impute NaN or missing value with median
4. Nominal encoder using one hot encoder

Customer segmentation

The machine learning pipeline for customer segmentation is consisted of:

1. Transformer to scale the feature using standardization
2. Principal component analysis estimator to reduce the dimensionality of the input features
3. Minibatch KMeans to cluster the data

Mailout response prediction

The machine learning pipeline for mailout response prediction is consisted of:

1. Transformer to scale the feature using standardization
2. Estimators to predict the mailout response, the candidate models are as follow:
    i. Logistic regression as the representative of the basic model
    ii. Random forest classifier as the representative of the bagging model
    iii. Xgboost classifier as the representative of the boosting model
3. The machine learning pipeline hyper parameters are tuned using BayesSearchCV.

Benchmark Model

Since there are no clear formulas to cluster the customer segmentation and to predict the mailout response, the benchmark model for customer segmentation and mailout response will be “majority guessing”. Meaning that, the models have to be better than just assigning the data with the majority label on a mailout response prediction task. The result of the model should satisfy the following criteria:

$$Model\ AUC\ ROC\ score\ \geq\ 0.5$$

Methodology

Implementation

The workflows for approaching the goals of this project are designed as follow:

1. Manually identify the null value
    a. From observing the “DIAS Information Levels - Attributes 2017.xlsx”, the null value for each feature in the demographic data is encoded differently. If it is not treated correctly, it will lead to confusion in data exploration and definitely pose a problem in the machine learning pipeline.
    b. The null value in this step will be saved into a JSON file for further used in the project.
2. Metadata exploration
    a. In this step, the description of each feature will be skimmed and filter out irrelevant or biased features (sex for example).
    b. The level of measurement will be identified in this step and save in JSON format.
3. Data exploration
    a. The individuals data will be explored, such as data distribution, data missing percentage or identifying quasi constant features. This information will be used in feature engineering steps to filter out unnecessary features or individuals to boost the performance of machine learning models.
4. Feature selection
    a. The input features are analyzed and select only some of it that are completed and contain meaningful information. The criterias being used in this project are as follow:
        i. Missing value percentage of each feature
        ii. Missing value percentage of each record
        iii. Correlation of each feature
        iv. Cardinality of each feature
        v. Quasi constant of each feature
5. Data cleaning
    a. To facilitate the modeling process, the data cleaning pipeline will be created such that raw pandas dataframe can be input to the pipeline. Thus, the data cleaning pipeline is consisted of:
        i. Column selecting transformer to select only the selected feature
        ii. Encode unknown value transformer to convert the missing value to NaN
        iii. Impute NaN or missing value with median
        iv. Nominal encoder using one hot encoder
6. Customer segmentation

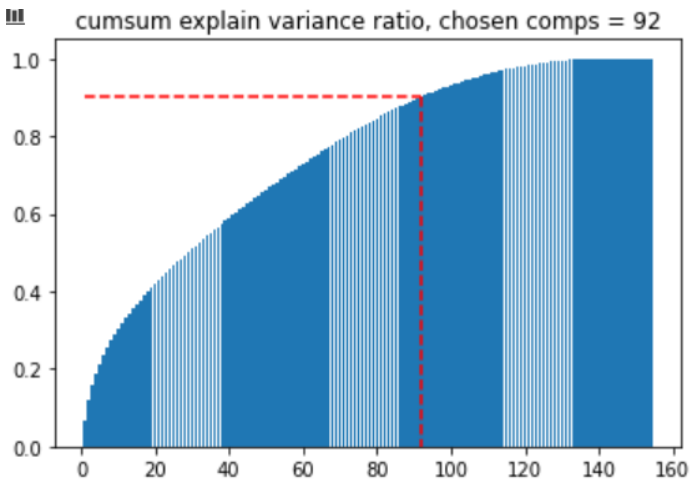
    a. *The appropriate number of PCA* components will be selected from the explained variance ratio. The retaining variance threshold is set at 0.9.
    b. *The appropriate cluster number of Mini Batch KMeans* will be calculated using elbow method.
    c. The machine learning pipeline with the two above parameters will be implemented to cluster the individuals in the “Udacity_AZDIAS_052018” and “Udacity_CUSTOMERS_052018” files.
    d. The results in this step will be helpful in the mailout response prediction task because the underlying information will be exploited to the supervised machine learning model in the next task.
7. Mailout response prediction
    a. The clustering result from customer segmentation will be appended to the existing feature.
    b. BayesSearchCV with logistic regression and l1 regularization will be quickly fit to the “Udacity_MAILOUT_052018_TRAIN” file.
    c. *Select only the most important feature* from the logistic regression parameters to reduce overfitting of the machine learning model. In this project, the top 20 absolute coefficients of all the features are selected.
    d. The BayesSearchCV and all of the candidate models will be fit to the data again with only the most important features identified from the above step.
8. Kaggle mailout response prediction
    a. Implement the fitted model pipeline in the previous step to predict the “Udacity_MAILOUT_052018_TEST” file
    b. Submit the result to kaggle website

Results

Customer segmentation

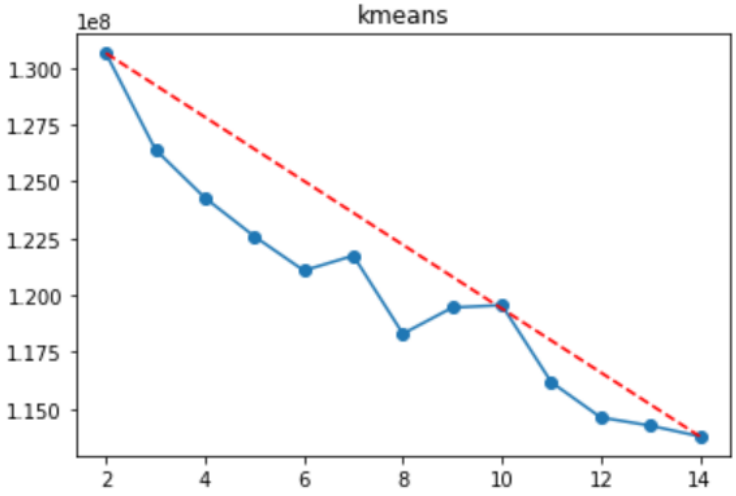
Select the appropriate number of PCA components

The cumulative sum of variance ratio can be shown in the below figure. The redline indicates the threshold value which is set at 0.9. The appropriate number of PCA components are 92 components.These components will, then, be used to identify the appropriate cluster number of Mini Batch KMeans.



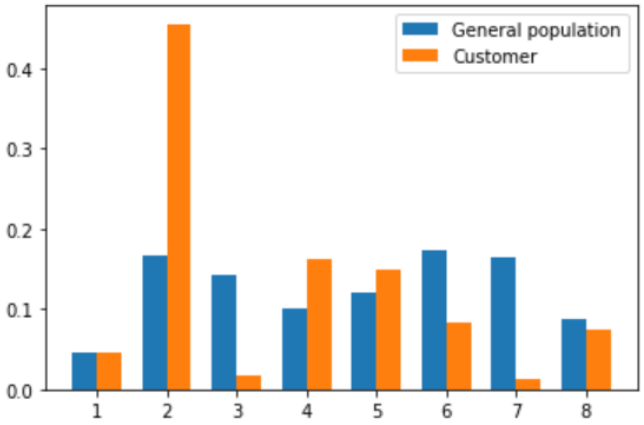
Select the appropriate number of Mini Batch KMeans

The cluster number of Mini Batch KMeans is ranging from 2 to 14. The inertia or sum of square distance at each cluster number is plotted as illustrated below. There are some inconsistencies in the curve from the local optima of Mini Batch KMeans, it is the trade-off for better calculation time. The estimated elbow point, or the point that has the most distance from the red line, is equal to 8.



Clustering Results

The above two parameters will be used and the whole clustering pipeline will be fitted again with the “Udacity_AZDIAS_052018” and “Udacity_CUSTOMERS_052018” data. The proportion of each cluster is illustrated below.



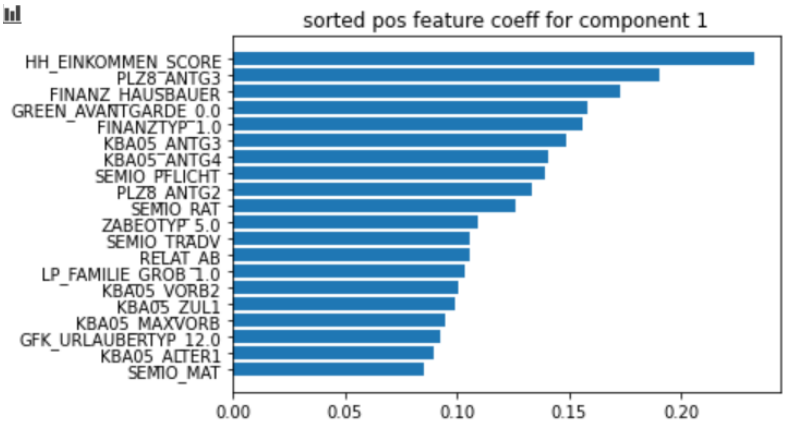
From the result, we can see that cluster 3 and cluster 7 are the group of individuals who are unlikely to be Arvato's customers. While cluster 2 is the group of individuals most likely to be Arvato's customers.

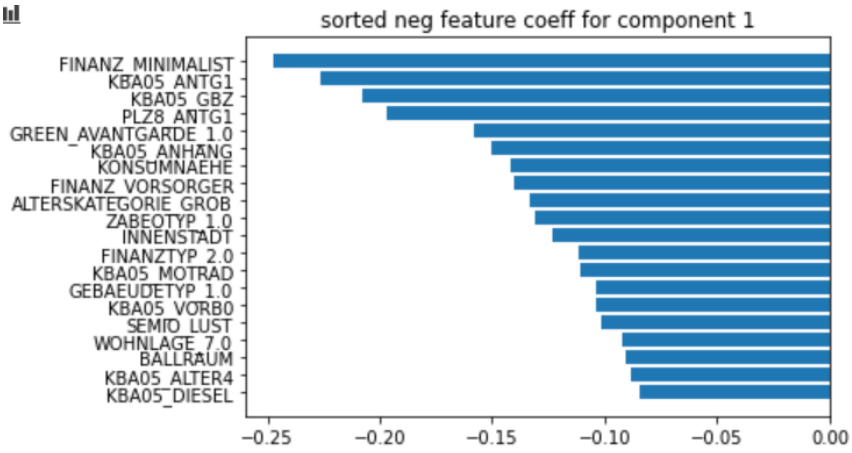
Since there are 92 components in the PCA model, it is impossible to explain all of its components. So, we will only extract the first components which contain the most explainable variance ratio as our representative of the group and the cluster 2, 3 and 7 which clearly distinguish the normal population and existing Arvato's customers.

The clustering center on the first component on cluster 2, 3 and 7 are shown in the table below. The cluster 2, the behavior of the individual who is likely to be Arvato's customer, has a negative coefficient on the first component. While cluster 3 and 7, the behavior of the individual who is unlikely to be Arvato's customer, has a positive coefficient on the first component.

	Cluster 2	Cluster 3	Cluster 7
comp_1	-3.275836	4.885325	2.060756

Since, the component of the PCA is just a projection from original features. We can calculate the coefficient of original feature on each PCA component to identify the characteristics of each component. The characteristic of the first PCA component, sorted by both negative and positive value, can be shown in the two below figures:

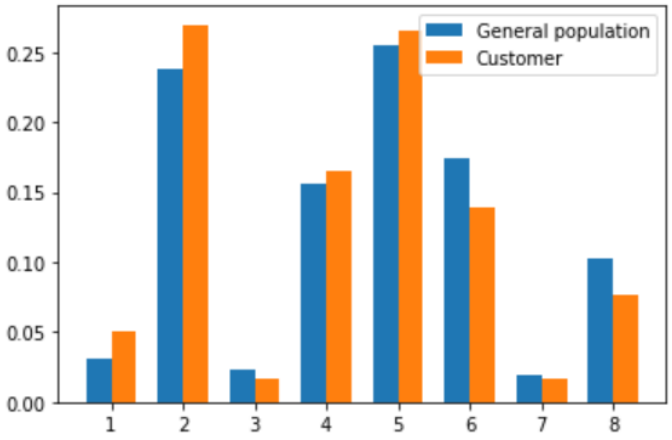




Mailout response prediction

Exploit Underlying Information

First, the underlying information of the “Udacity_MAILOUT_052018_TRAIN” will be exploited using the clustering pipeline from the previous task. The result of the clustering can be shown below:



The clustering result may not clearly distinguish the response individual and un-response individual, since the behavior of the response individual may be different from the existing customers. However, this information will be useful to the mailout prediction model and add to the input as additional feature.

Select Most Important Features

BayesSearchCV with the objective to maximize AUC ROC and logistic regression is quickly implemented to find the top 20 important features as shown in the below table:

feature	coeff
LP_FAMILIE_GROB_0.0	-0.361583
KBA05_VORB0	0.266885
FINANZ_VORSORGER	0.178453
KBA05_DIESEL	0.173950
LP_FAMILIE_GROB_1.0	0.154031
INNENSTADT	-0.153707
HH_EINKOMMEN_SCORE	-0.151675
KBA05_ZUL4	0.138644
SEMIO_ERL	-0.138307
SEMIO_VERT	-0.137523
LP_FAMILIE_GROB_3.0	0.133335
KBA05_MOD2	0.126920
GEBAEUDETYP_6.0	0.126852
SEMIO_TRADV	-0.124140
SEMIO_LUST	0.119296
WOHNLAGE_0.0	-0.117795
CLUSTER_0.0	**0.110420**
MIN_GEBAEUDEJAHR	0.108788
SEMIO_SOZ	0.105854
BALLRAUM	0.102822

Note that the CLUSTER_0.0 feature (cluster number 1) is selected as one of the most important features since it can distinguish the response and un-response behavior of the individuals in the data.

Prediction Results

The whole pipeline with BayesSearchCV of all the candidate models will be fitted with the most important features identified in the previous step. The result can be show in the below table:

Model	Best CV score	Refit training score
Logistic regression	0.579	0.693
Random forest classifier	0.594	0.686
XGboost classifier	0.568	0.643

From the above result, the most promising model is the Random forest classifier.

Conclusions

Customer segmentation

From customer segmentation results, the clustering centers coefficient of the first component can be map to the characteristic of the positive and the negative original feature coefficient of the first component to identify the characteristic of each cluster group.

The cluster 2, representative of the existing customers tends to have more value on FINANZ_MINIMALIST. According to the “DIAS Attributes - Values 2017.xlsx”, the existing Arvato’s customers are the individuals who have low financial interest.

While the cluster 3 and 7, representative of the individuals who are unlikely to be Arvato’s customers, tends to have more value on HH_EINKOMMEN_SCORE which indicates the low income individuals.

In summary, the Arvato’s customer characteristics are high income with low financial interest.

Mailout response prediction

From mailout response prediction, even though all three models can surpass the benchmark of 0.5 AUC ROC score. However, There are room of improvement indicating by the following signs:

1. There is still a room to improve the model complexity, since refit training score is far lower than 1.
2. The difference between CV score and refit training score are high. It is the sign of the overfitting,

Thus, the model can be improved by the following finetune methodologies.:

1. Better finetune the hyperparameters with longer time and wider range to increase both model complexity and finding the optimize regularization
2. Improve feature engineering process by carefully analyze each feature, discussing with field's expert to manually choosing the feature and the imputation method on each feature

However, we have to consider the trade-off between time-to-market of the product and squeezing the most performance of the model. Note that, the as is models are good enough to be implemented as Minimum Viable Product since it surpasses the benchmark model.

References

[1] "Arvato - Bertelsmann SE & Co. KGaA," Bertelsmann, [Online]. Available: https://www.bertelsmann.com/divisions/arvato/#st-1. [Accessed 15 03 2021].

[2] P. Premkanth, "Market Segmentation and Its Impact on Customer Satisfaction with Especial Reference to Commercial Bank of Ceylon PLC," Global Journal of Management and Business Research, pp. 33-40, 2012.

[3] "Customers Insight," AZ Direct, [Online]. Available: https://www.az-direct.com/site/en/products/customer-insights/. [Accessed 15 03 2021].

[4] "Classification: Prediction Bias," Google, [Online]. Available:https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc#:~:text=AUC%3A%20Area%20Under%20the%20ROC,to%20(1%2C1) [Accessed 09 05 2021].

[5] Tabachnick, B. G., & Fidell, L. S. (1996). Using Multivariate Statistics (3rd ed.). New York: Harper Collins.