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

The main goal of this project is to uplift the response of the Arvato's mailout by inventing the machine learning pipeline to predict the probability of the response of the new customer facilitating the business team to approach the customer effectively. The general individual demographic data from AZ DIAS database [3], existing Arvato's customer demographic data and historical Arvato's mailout response data will be explored and utilized to achieve this goal.

The solution is consisted of 2 consecutive tasks as follow:

1. Customer segmentation

Exploit the latent information from demographic data by clustering the individuals with similar characteristics to the same group. The potential customer characteristic will be identified and fed to the next task.

2. Mailout's response prediction

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

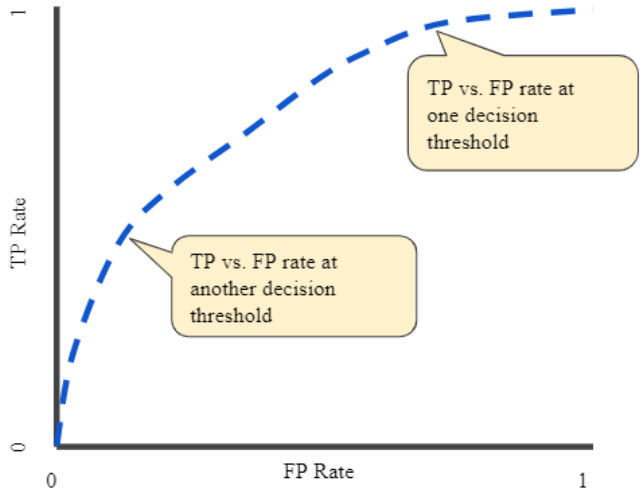
The approach to the above 2 tasks can be divided into 7 steps as follow:

1. Manually explore the metadata provided by Arvato's
    a. The null value in each feature is not left as blank and encoded differently. Thus, it is needed to be encoded as null and imputed appropriately.
    b. The level of measurement of each feature is not explicitly clarified. It needs to be labeled manually and processed accordingly as one of the following levels.
        i. Nominal categorical value
        ii. Ordinal categorical value
        iii. Interval value
        iv. Temporal value
    c. The two above additional metadata will be saved into a JSON file for further used in the project.
2. Exploratory Data Analysis (EDA)
    a. The demographic data will be thoroughly explored, such as
        i. Label imbalancement
        ii. Feature missing percentage to filter out sparse feature
        iii. Record missing percentage to filter out sparse record
        iv. Feature correlation
        v. Quasi constant feature
        vi. Cardinality feature
    b. The above EDA will be used in feature engineering steps to filter out unnecessary features or records to boost the performance of the machine learning pipeline.
3. Feature selection
    a. The criteria will be set according to the EDA processes to automatically identify the meaningful features and records.
4. Data cleaning
    a. To facilitate the modeling process in automatic fashion, the data cleaning pipeline will be created such that raw pandas dataframe can be fed into. The data cleaning pipeline is consisted of:
        i. Column selecting transformer to select only the subset of the feature identifying in the feature selection step

    ii. Encode unknown value transformer to convert the encoded missing value to nan using the metadata in the first step
    iii. Imputing transformer to impute nan value with median of the remaining data
    iv. Nominal encode transformer to encode nominal value using one hot encoder
5. Customer segmentation
    a. The cleaned demographic data will be used to cluster the individuals with same characteristics with the following 3 subtasks
        i. Normalize the input data with standardization to levelized the standard deviation which is greatly impact the Principal Component Analysis
        ii. Reduce the number of dimensionality using PCA
        iii. Cluster the individuals from the PCA components
    b. Identify the characteristic of each clustered group from the PCA component's coefficient which can be map to the original features
6. Mailout response prediction
    a. The clustering result from the fifth steps will be used as the additional features together with the demographic data from the historical Arvato's mailout data.
    b. Since there are more around 360 features, to prevent overfitting, only a handful of the most important feature will be quickly selected from the following pipeline
        i. Normalization using standardization
        ii. Logistic regression with elasticnet regularization
        iii. Hyperparameters tuning the above classification model using BayesSearchCV
        iv. Select a handful of features from the top most absolute coefficient
    c. The selected features will be used to fit the following mailout response prediction pipeline
        i. Normalization using standardization
        ii. Classification machine learning model with the following candidate models
            1. Random forest classifier as the representative of the bagging model
            2. XGboost classifier as the representative of the boosting model
        iii. Hyperparameters tuning the above classification model using BayesSearchCV
7. Kaggle mailout response prediction
    a. All of the above information and pipelines which consist of data cleaning pipeline, customer segmentation pipeline and mailout response prediction pipeline will be implemented to "Udacity_MAILOUT_052018_TEST" demographic data and submit the result to the kaggle website to identify the potential of solution

Evaluation Metrics

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 solution. 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:

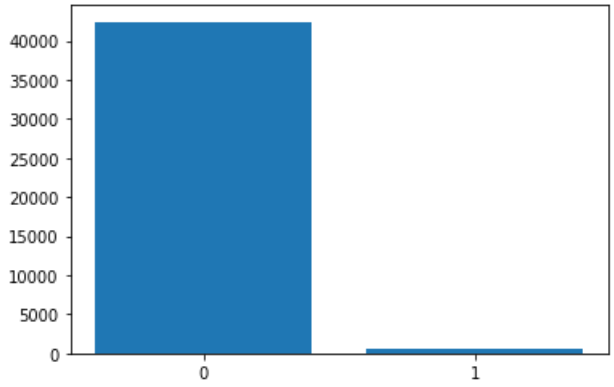
1. DIAS Attributes - Values 2017.xlsx
    a. This file contains all possible values and how null value is encoded for each feature.
2. DIAS Information Levels - Attributes 2017.xlsx
    a. This file contains the description of each feature.

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.

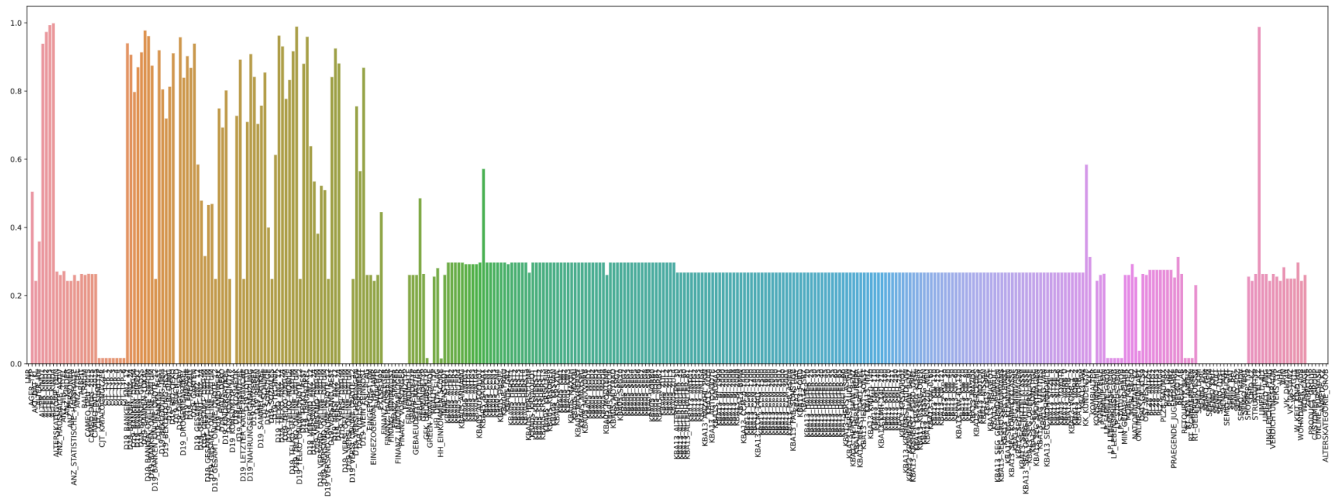
Feature missing percentage observing

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.

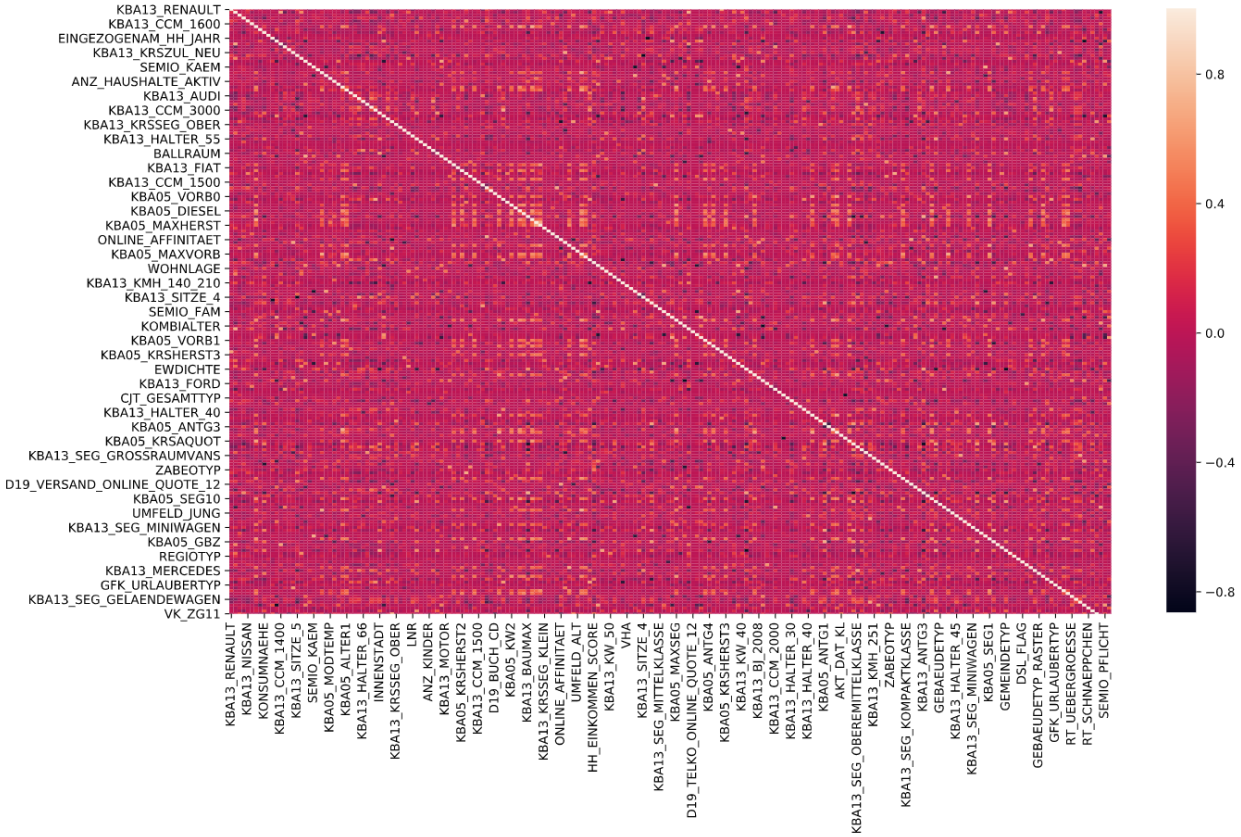
By setting threshold of 30%, the percentage of the record that exceed the threshold can be shown below:

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

In this section, the algorithms and machine learning techniques for main tasks which are Customer Segmentation and Mailout Response Prediction will be described.

Customer Segmentation

Customer Segmentation purpose is to cluster the customer with similar characteristics to the same group. The pipeline to effectively approach this goal is composed of:

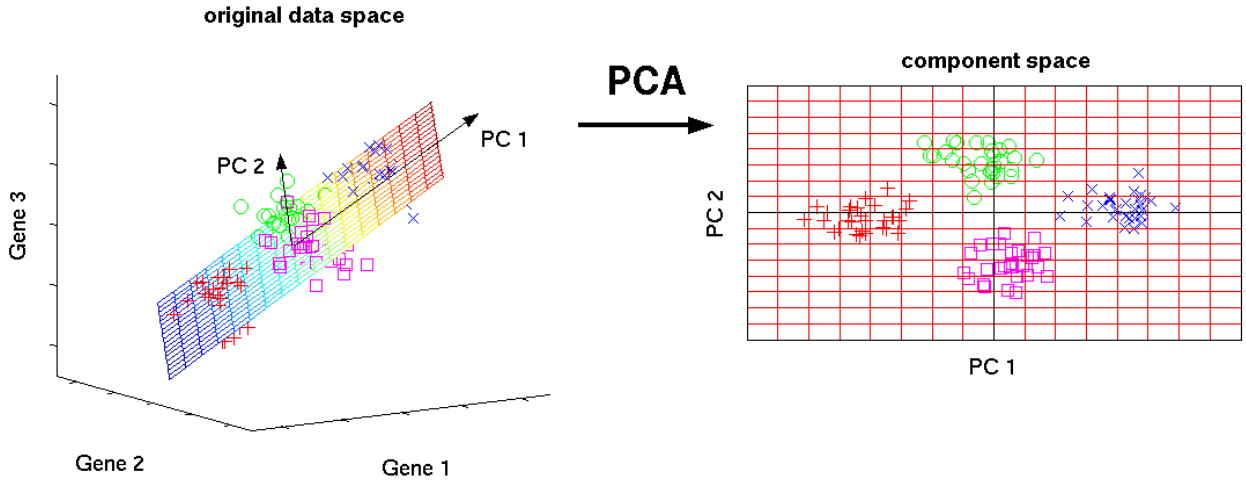
1. Normalization

The input features magnitude will be normalized using standardization to levelized the standard deviation and mean of the features to 1 and 0 respectively. This characteristic will reduce the feature bias of the Principal Component Analysis from different scales on feature's variance.

2. Principal Component Analysis (PCA)

Since the distance of the clustering models, especially KMeans clustering, does not qualitatively meaningful on high dimensional features [6]. The concept of far distance and near distance, which is intuitively meaningful in low dimensional, will quantitatively lower with the more number of dimensions. Meaning that, the individuals' characteristics in the same cluster are not necessarily distinct from another cluster. This effect is called "Curse of Dimensionality".

Thus, PCA will be implemented to reduce the dimension of the feature while retaining the information as much as possible. The PCA objective function is to project the data from high dimension space to a new coordinate system with lower dimension.



However, the more dimension is reduced, the more variance or meaningful information will vanish. Hence, the criteria in choosing the number of components is needed. In this project, the number of components is chosen as minimum as possible while still retaining 90% of the original variance.

3. Mini Batch KMeans Clustering.

Mini Batch KMeans is chosen in this project since its training time is lower than original KMeans clustering resulting in a shorter research cycle in improving time to market of the model. The concept of Mini Batch KMeans is similar to original KMeans except that the Mini Batch KMeans will update its cluster on a small subset of the data (batch) instead of the whole data set like original KMeans.

The time complexity of original KMeans clustering is $O(n^2)$ [7] where n is the number of the record. By dividing data set to smaller subset, for instance k fold, the time complexity will be

$$O\left(\left(\frac{n}{k}\right)^2 \times k\right) = O\left(\frac{n^2}{k}\right)$$

From the above equation, by increasing k, the computation time will be lower. And this methodology of dividing the data set into mini batches before KMeans clustering is called Mini Batch KMeans clustering. However, the probability that the optimization will be stuck in local minima is also increased following the increasing number of k, since the model is only fit on the mini batch at a time.

Furthermore, choosing the appropriate number of clusters is another integral part in Customer Segmentation. In this project, the elbow method will be implemented as a criteria in choosing the number of clusters.

Mailout Response Prediction

Mailout Response Prediction purpose is to predict the probability that the individuals will respond to the Arvato's campaign email. This model will screen out individuals who are unlikely to be Arvato's customer even before approaching them. The pipeline to effectively approach this goal is composed of

1. Customer Segmentation

Customer Segmentation result is an integral part in the Mailout Response Prediction task. By utilizing the latent information learned from demographic data of the general population and existing Arvato's customers will boost the performance of the Mailout Response Prediction task.

2. Normalization using standardization to reduce the bias of greater magnitude feature
3. Classification Model

There are 3 candidates of the classification models which are Logistic Regression, Random Forest Classification and XGBoost classification.

a. Logistic Regression is the representative of the most basic model in this project. The objective of the Logistic Regression is to minimize the log loss function where $\widehat{y}$ is the prediction probability and y is the ground truth label.

$$log\ p(y|x) = y\ log\widehat{y} + (1 - y)\ log\ (1 - \widehat{y})$$

b. Random Forest Classification is the representative of the bagging ensemble model in this project. The random forest is the ensemble of many decision trees which the result of the classification is derived from hard voting of these decision trees.
The common objective function for decision trees in random forest is to minimize Gini Impurity where p_k is the proportion of the k^{th} class of the label. The more Gini Impurity means the more mixed of the label in any particular node.

$$G = \sum_{k=1}^{K} p_k(1 - p_k)$$

c. XGBoost classification is the representative of the boosting ensemble model in this project. XGBoost is based on the Gradient Boosting concept but instead of fitting on first order gradients like Gradient Descent optimization technique, the XGBoost is fitted with the parameters on second order derivatives, called Hessian more like Newton's optimization technique.
However, the objective function for binary classification is the same as logistics regression which is log loss function.

4. BayesSearchCV

The continuous optimization technique derived from the exact mathematical formula of the problem can only fit the parameters of the classification model but not the hyperparameters, because if it is included in the problem, the mathematical formula will not be able to formulate. The hyperparameters include regularization coefficient in logistic regression, number of decision trees in Random Forest Classifier and XGBoost etc.

There are various techniques in tuning these hyperparameters such as Random search on the given space or Grid search on the equal interval on the given space. In this project, the Bayesian optimization, the blackbox optimization technique, will be implemented. The blackbox optimization estimates the mathematical formula and objective function instead of deriving the exact formula which in some cases is impossible. Surrogate model is the common implementation to estimate the exact formula of the problems [8].

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

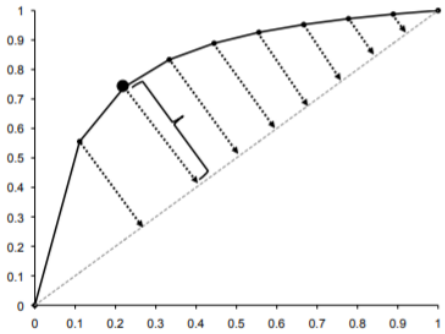
Implementation

In this section, the details of the criteria and configuration of the approach being used in the project will be described.

1. Manually explore the metadata provided by Arvato's

a. The null encoded value in each feature was manually read from given “DIAS Information Levels - Attributes 2017.xlsx” and saved in JSON format feature by feature.
b. The level of measurement on each feature was, also, read from given “DIAS Information Levels - Attributes 2017.xlsx” and saved in JSON format feature by feature, because different levels of measurement require different preprocessing, especially nominal categorical feature.

2. Exploratory Data Analysis (EDA)
    a. The demographic data was thoroughly explored, such as
        i. Label imbalancement
        ii. Feature missing percentage to filter out sparse feature
        iii. Record missing percentage to filter out sparse record
        iv. Feature correlation
        v. Quasi constant feature
        vi. Cardinality feature
3. Feature selection
    a. The input features were analyzed and selected only the features that contain clean and meaningful information using pre-set criteria as follows.
        i. Missing value percentage of the feature is less than 30%.
        ii. Missing value percentage of each record is less than 30%.
        iii. Correlation of any feature on every other feature is less than 0.7.
        iv. Cardinality of the feature or class of the value of the nominal feature is less than 20.
        v. The feature is not quasi constant or the mode of the value in the feature is less than 90% of the total size.
4. Data cleaning
    a. Data cleaning pipeline details are as follow:
        i. Select the features that passing all the criteria in the above steps and save as customize sklearn component
        ii. Convert the encoded missing value to nan using the metadata in the first step to facilitate the imputing step
        iii. Imputing missing value with median
        iv. Convert nominal feature with N distinct value to N additional features
5. Customer segmentation
    a. The number of components was chosen as minimum as possible while still retaining 90% of the original variance.
    b. The elbow method was implemented as a criteria in choosing the number of clusters. The Kneedle algorithm [9] is used to facilitate the decision as shown in the following figure.



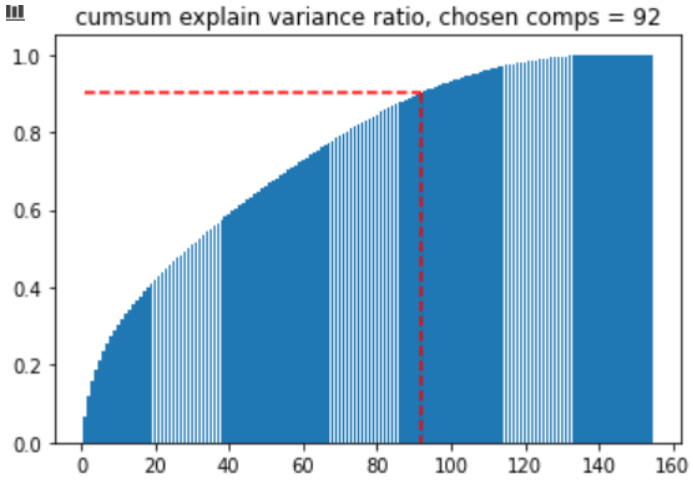
6. Mailout response prediction
    a. In the first step of this task, the Customer Segmentation pipeline was used to cluster the individuals in Udacity_MAILOUT_052018_TEST.csv file.
    b. The clustering result was appended to the original demographic data in Udacity_MAILOUT_052018_TEST.csv file.
    c. The logistic regression with BayesSearchCV was quickly implemented to find the most important features.
    d. The most important feature was chosen from the top 20 coefficient parameters of the logistic regression model.
    e. The selected most important features were used to fit BayesSearchCV with Random Forest Classifier and XGBoost classifier.
    f. The hyperparameters used for tuning each classification model are as follow:
        i. Logistic regression:
            1. C (regularization coefficient)
            2. L1 Ratio (the ratio between L1 and L2 regularization)
        ii. Random Forest Classifier:
            1. N Estimators (number of decision trees in the forest)
            2. Max depth of each decision trees
            3. Min sample split, the minimum number of samples to split any node
            4. Min sample leaf or the minimum number of sample required to be in any leaf node after splitting
        iii. XGBoost
            1. N Estimators (number of decision trees in the forest)
            2. Learning rates, the coefficient of the step to update parameters in XGBoost
            3. Max depth of each trees
            4. Scale for positive weight to deal with imbalance label

Results

Customer segmentation

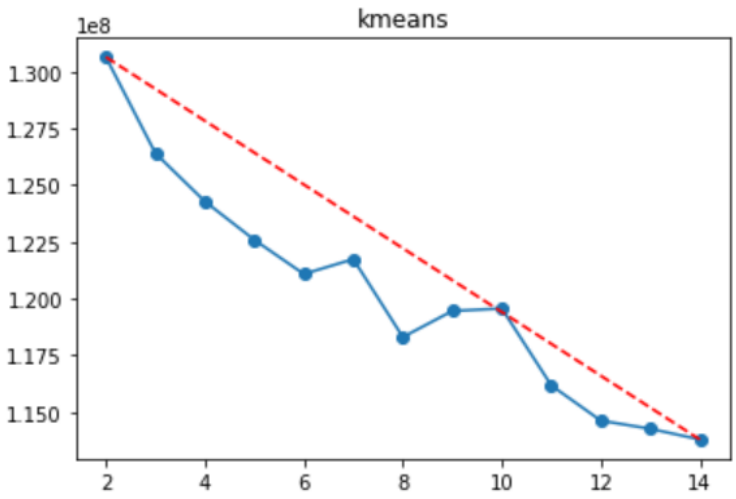
Select the appropriate number of PCA components

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.



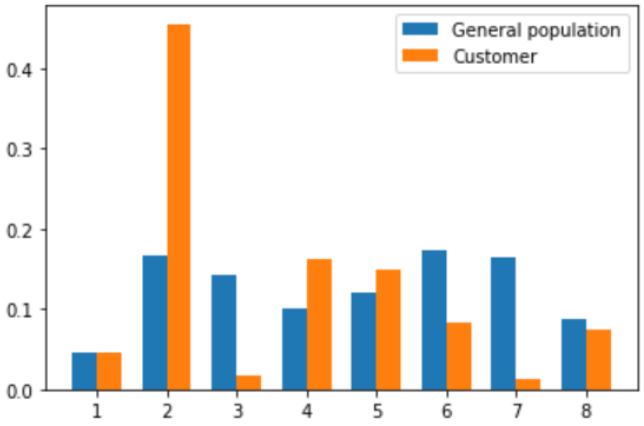
Select the appropriate number of Mini Batch KMeans

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.



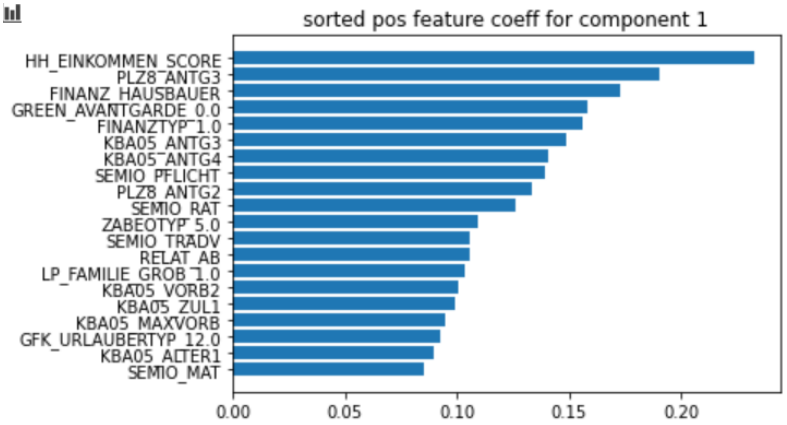
From the result, we can see that cluster 3 and cluster 7 are the group of individuals who are unlikely to be Arvato’s customers. While cluster 2 is the group of individuals most likely to be Arvato’s customers.

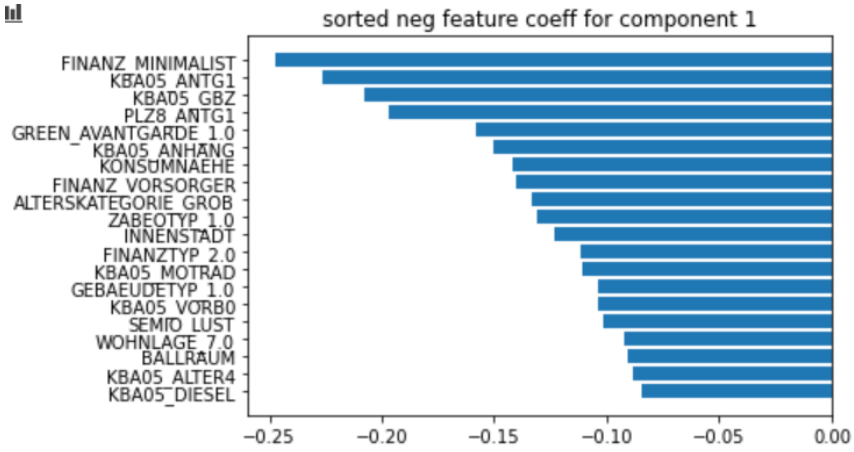
Since there are 92 components in the PCA model, it is impossible to explain all of its components. So, we will only extract the first components which contain the most explainable variance ratio as our representative of the group and the cluster 2, 3 and 7 which clearly distinguish the normal population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

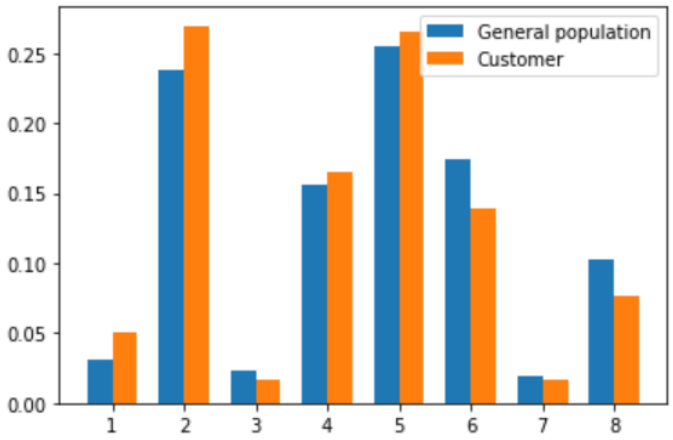




Mailout response prediction

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.

[6] Aggarwal C.C., Hinneburg A., Keim D.A. (2001) On the Surprising Behavior of Distance Metrics in High Dimensional Space. In: Van den Bussche J., Vianu V. (eds) Database Theory — ICDT 2001. ICDT 2001. Lecture Notes in Computer Science, vol 1973. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-44503-X_27

[7] Pakhira, Malay. (2014). A Linear Time-Complexity k-Means Algorithm Using Cluster Shifting. Proceedings - 2014 6th International Conference on Computational Intelligence and Communication Networks, CICN 2014. 10.1109/CICN.2014.220.

[8] "The intuitions behind Bayesian Optimization with Gaussian Processes," Charles Brecque, [Online]. Available: https://towardsdatascience.com/the-intuitions-behind-bayesian-optimization-with-gaussian-processes-7e00fcc898a0. [Accessed 11 05 2021].

[9] V. Satopaa, J. Albrecht, D. Irwin and B. Raghavan, "Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior," 2011 31st International Conference on Distributed Computing Systems Workshops, 2011, pp. 166-171, doi: 10.1109/ICDCSW.2011.20.