**REPORT FOR UDACITY-ARVATO CUSTOMER SEGMENT PROJECT**

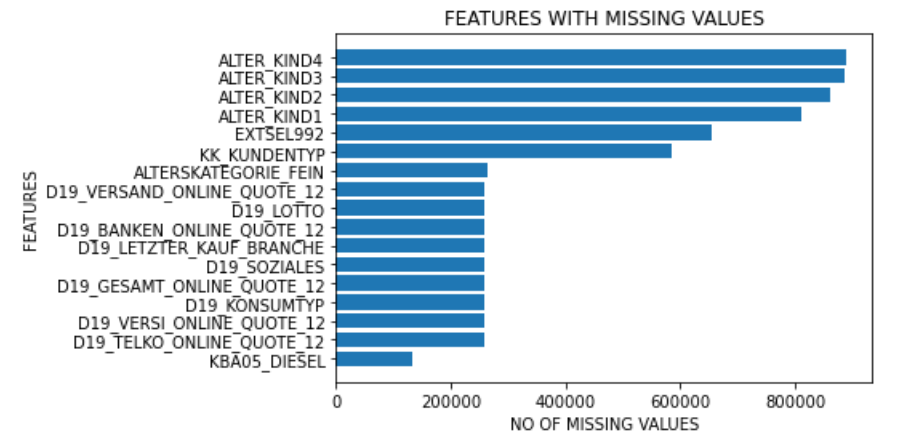
DROPPING FEATURES WITH TOO MANY MISSING VALUES

First we will check for features with missing values.

At first glance, removing features with 257113 missing values or more seems prudent. However, when I got to the 3rd part of this project, I realised D19\_SOZIALES is the most important feature in predicting potential customers of a mailout campaign. As such, I decided to only remove features with 300000 missing values and above. Therefore the following features are removed:

ALTER\_KIND4, ALTER\_KIND3, ALTER\_KIND2, ALTER\_KIND1, EXTSEL992 and KK\_KUNDENTYP.





Removing features with too many missing values stage 1

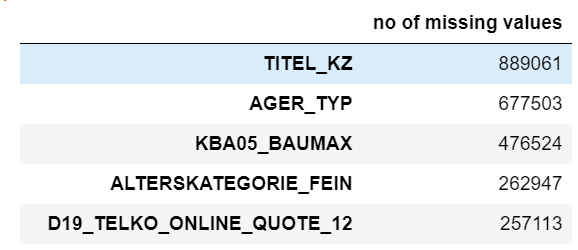
Upon inspecting the DIAS Attributes - Values 2017 document, I found features that have values that are labelled unknown. Therefore unknown values are similar to missing values and have to be converted to NaN. Therefore I extracted features that have unknown values. There are features that have different names in the azdias file namely CAMEO\_DEUINTL\_2015, KBA13\_CCM\_1400\_2500 and SOHO\_FLAG which are renamed to CAMEO\_INTL\_2015 , KBA13\_CCM\_1401\_2500 and SOHO\_KZ.

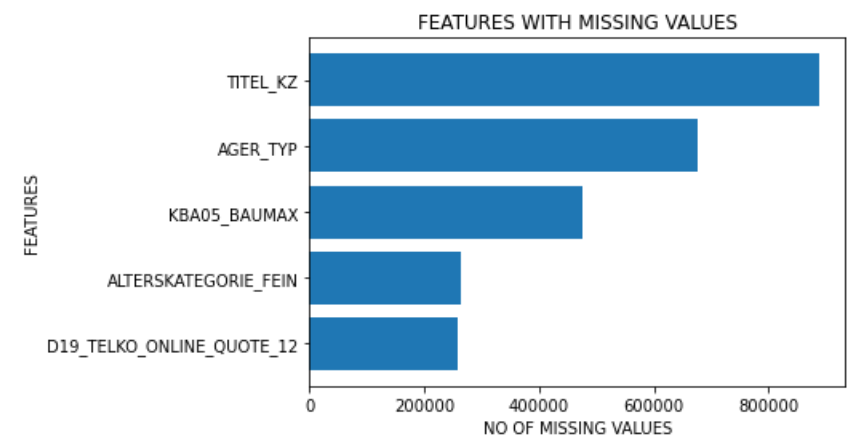
|  |  |
| --- | --- |
| -1, 9 | unknown |

|  |  |
| --- | --- |
| -1, 0 | unknown |

There are also features that are not in the AZDIAS dataframe, namely BIP\_FLAG, D19\_KK\_KUNDENTYP, GEOSCORE\_KLS7, HAUSHALTSSTRUKTUR , WACHSTUMSGEBIET\_NB. There are also 85 features in AZDIAS that are not in the DIAS Attributes document. Therefore for these 85 features, I am unable to convert unknown values to NaN since I do not know which values will be labelled unknown.

After converting unknown values to NaN, I check for features with missing values again.





Removing features with too many missing values stage 2

Since TITEL\_KZ, AGER\_TYP AND KBA05\_BAUMAX exceed 400000 missing values, they are removed accordingly.

FURTHER DATA CLEANING AND FEATURES PREPROCESSING

LP\_STATUS\_GROB and LP\_FAMILIE\_GROB are dropped because they are similar to LP\_STATUS\_FEIN and LP\_FAMILIE\_FEIN, the latter two having more granular disaggregations. LNR is dropped because it is an identification number list. CAMEO\_DEU\_2015 and LP\_LEBENSPHASE\_FEIN are dropped because they are categorical and nominal features and they have 44 and 40 categories respectively. It is unwise to convert to dummy variables which will lead to sparse matrix, therefore the decision is to drop them. D19\_LETZTER\_KAUF\_BRANCHE is dropped because it is a duplicated information.

OST\_WEST\_KZ and ANREDE\_KZ are binary features with string values and therefore the values are converted into 0 and 1.

FEATURES ENGINEERING

For PLZ8\_BAUMAX, I created a new feature that indicates if someone is living in a business building or not. For PRAEGENDE\_JUGENDJAHRE, a mainstream or avant-garde movement feature and another feature grouping together similar decades are created. For WOHNLAGE, a new feature indicating if someone live in a rural or urban region is created. For CAMEO\_INTL\_2015, it is separated into a feature indicating household wealth status and a feature indicating family types. EINGEFUEGT\_AM feature is in a string format, eg: 3/11/1993 0:00. Therefore the dates are extracted and converted into months. I used 1/1/2019 as the end date.

Next we will have to fill NaN values with either mean values for continuous features or mode for ordinal/nominal features. We identified EINGEZOGENAM\_HH\_JAHR , MIN\_GEBAEUDEJAHR, KBA13\_ANZAHL\_PKW, ANZ\_HAUSHALTE\_AKTIV, VERDICHTUNGSRAUM and ANZ\_STATISTISCHE\_HAUSHALTE as continuous features.

SELECTING MOST IMPORTANT FEATURES FOR KMEANS ANALYSIS

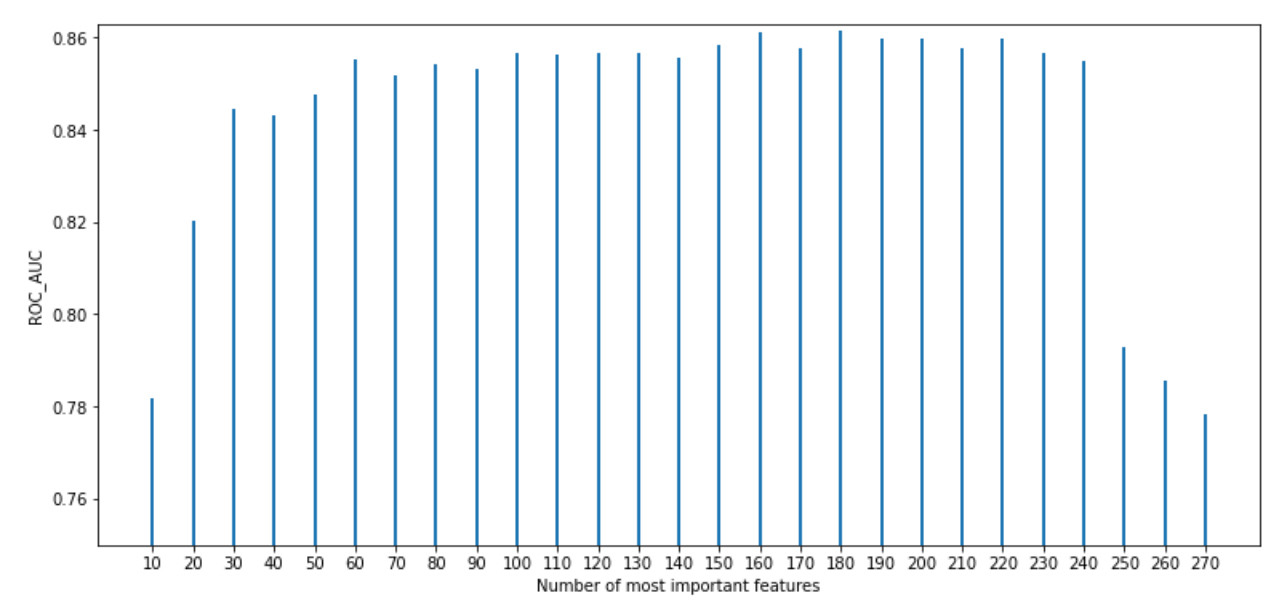
Once the NaN values are filled, I was considering Principal Component Analysis, a dimensionality reduction technique that can convert features into principal components that account for most of data variance. While PCA can deal with correlated features, however it can’t deal with redundant features that have little value in predicting potential customers of the mail order company. While I did performed PCA initially, when I completed part 3 using XGBoost Classifier, I could extract the most important features in predicting potential customers. When I uploaded my predictions for MAILOUT TEST to Kaggle, my predictions earned a ROC-AUC score of 0.88149, placing me first in the leaderboard. I arbitrarily selected the 130 most important features in training my model.

What I did was first training my XGBoost model on the MAILOUT TRAIN and optimising hyperparameters using Bayesian Optimisation. I then test my model on MAILOUT TEST and I selected the model that earned me the best possible score in Kaggle, an AUC of 0.803. Then using only 130 most important features, I trained my model on both AZDIAS and CUSTOMER dataset and MAILOUT TRAIN is used as validation dataset. Using the best hyperparameters combination, my model earned me top spot in leaderboard with AUC 0.88149.

I would like to find out if there is an optimal number of important features for predicting potential customers. Therefore using the best hyperparameters combination, now I can retrain my model. I started training with 10 most important features and loop through additional ten features till I reached 270 features. Following are the ROC\_AUC scores when I used my trained model to predict on MAILOUT TRAIN.

(Format is No Of Features: ROC\_AUC score)

{10: 0.7818, 20: 0.8202, 30: 0.8445, 40: 0.8429, 50: 0.8477, 60: 0.8553, 70: 0.8519, 80: 0.8541, 90: 0.8531, 100: 0.8567, 110: 0.8565, 120: 0.8566, 130: 0.8567, 140: 0.8555, 150: 0.8583, 160: 0.8610, 170: 0.8575, 180: 0.8615, 190: 0.8598, 200: 0.8599, 210: 0.8578, 220: 0.8596, 230: 0.8566, 240: 0.8549, 250: 0.7928, 260: 0.7855, 270: 0.7782}



Determining the optimal number of most important features for K-means analysis

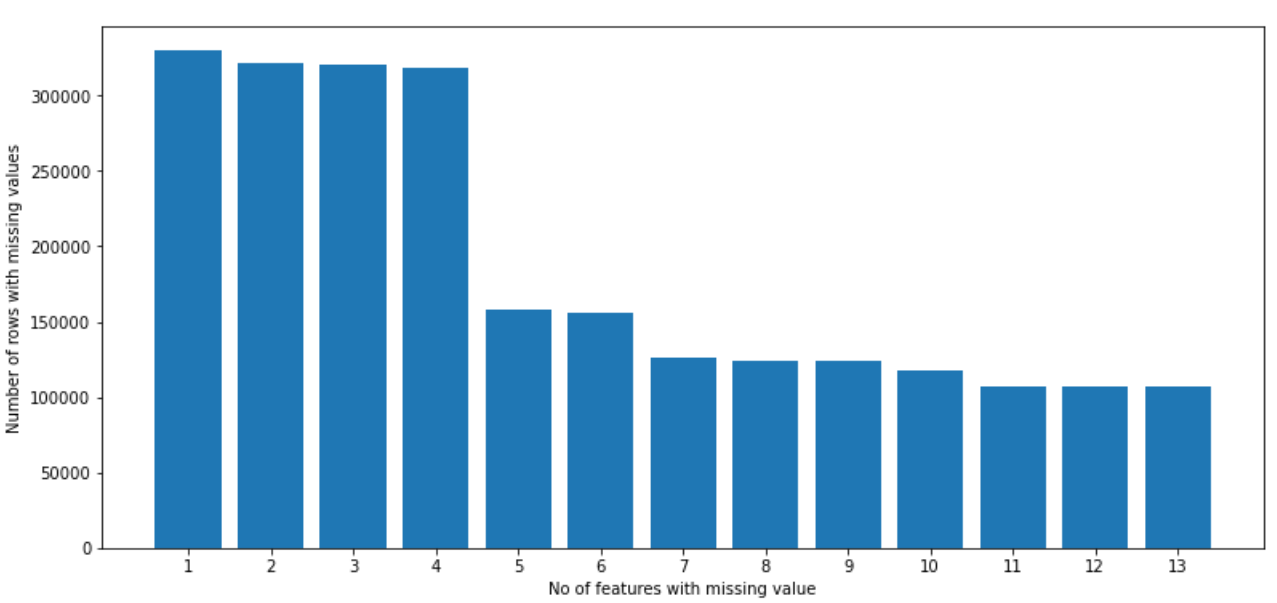
As can be seen above, the AUC score improved dramatically from 10 to 30 features. AUC scores continue to increase gradually until 60 features. Therefore I believe 60 features are the most optimal. If intrepretability of features is a concern of the mail order company, 30 features seem sufficient to train a robust predictive models. Therefore, I chose to eschew PCA and instead would perform customer segmentation using these 60 most important features.

REMOVING ROWS WITH TOO MANY MISSING VALUES

Before that I would need to remove rows with too many missing values.

{ 1: 329826, 2: 321895, 3: 320564, 4: 318123, 5: 157664, 6: 155907, 7: 125949,

8: 124394, 9: 124203, 10: 117147, 11: 107018, 12: 106747, 13: 106575}



Removing rows with too many missing values

Removing rows with 5 or more missing values seem reasonable without sacrificing too many examples. I could instead remove 10 or more missing values with many fewer examples discarded, however since I only retained 60 most important features, 10 is a tall proposition.

K-MEAN ANALYSIS

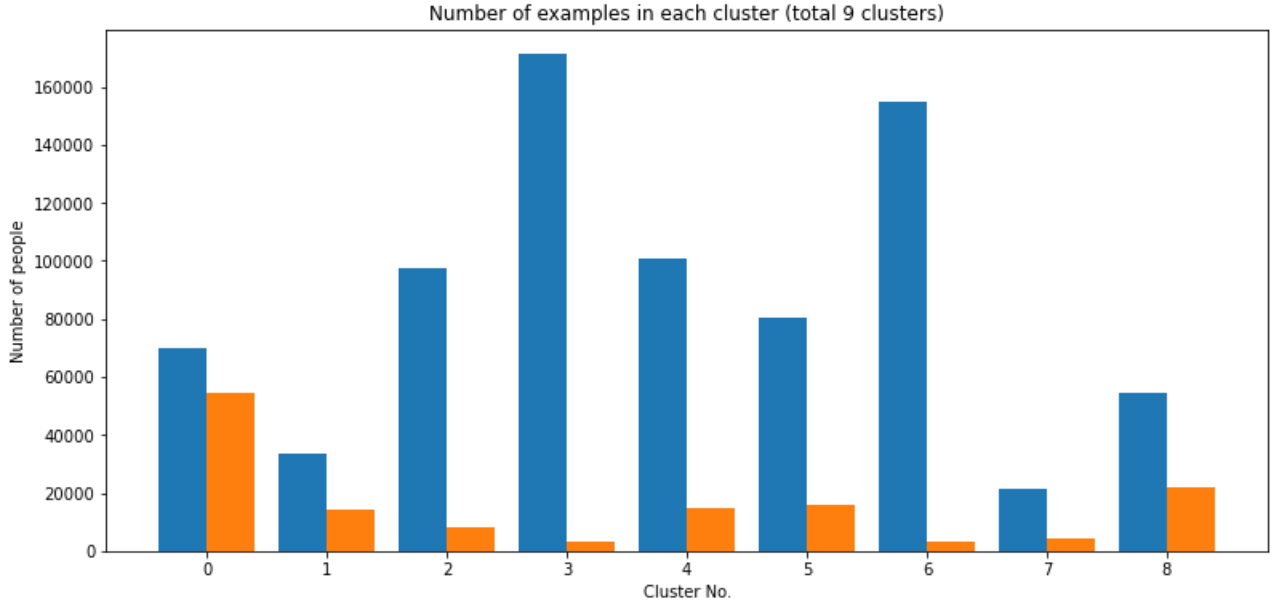
To select the number of clusters for K-Means analysis, I also chose to eschew the conventional method, the elbow method. This is because the aim of the customer segmentation differs somehow from the objective of elbow method. I would instead conduct trial and errors starting from 4 clusters until 15 clusters. The optimal number of cluster I choose will depend on if a particular cluster contains a significant proportion of customer. 9 clusters is selected based on this criteria.

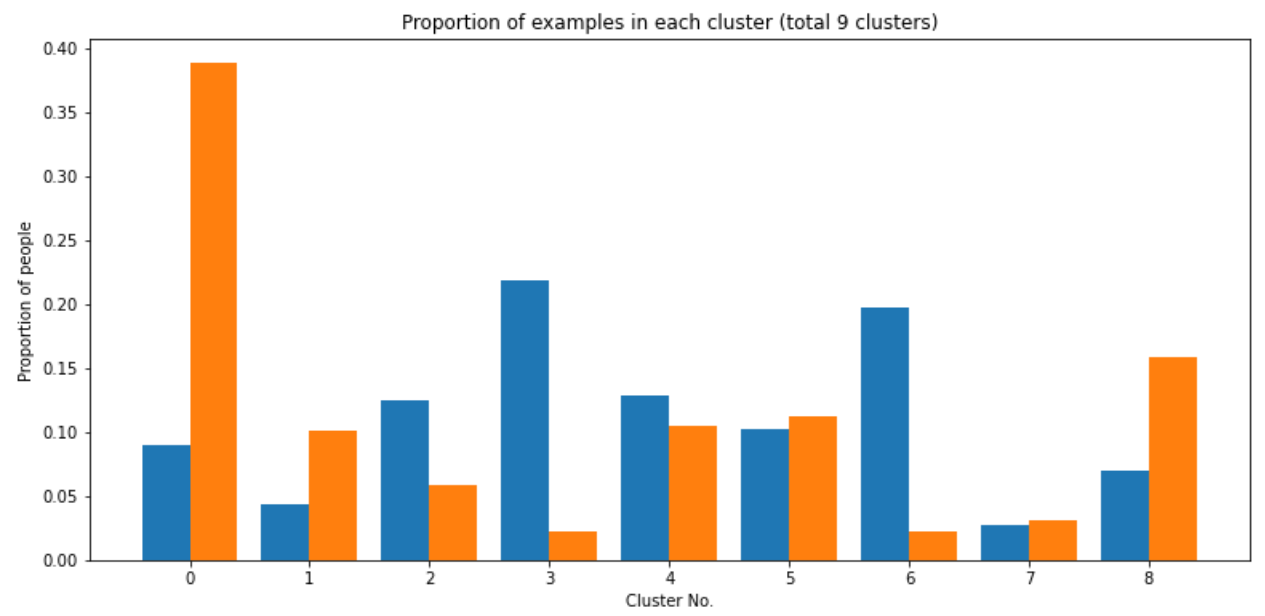
{**0: 70103**, 1: 33633, 2: 97520, 3: 171205, 4: 100766, 5: 80193, 6: 155008, 7: 21552, 8: 54530}

{**0: 54491**, 1: 14133, 2: 8278, 3: 3175, 4: 14689, 5: 15772, 6: 3207, 7: 4414, 8: 22195}

As can be seen, cluster 0 captures almost 40% of total customers of the company.

Potential customer in AZDIAS are 70103 people, representing 8.9% of population, a reasonable proportion.





Determining cluster that contains potential customers

For completeness of analysis, I also conducted K-Means analysis using 30 and 45 features.

30 features:

{0: 52102, 1: 219290, **2: 82996**, 3: 62791, 4: 204169, 5: 50856, 6: 33172, 7: 36094, 8: 43040}

{0: 6870, 1: 5186, **2: 56832**, 3: 17474, 4: 9075, 5: 13226, 6: 5654, 7: 6944, 8: 19093}

45 features:

{0: 154900, 1: 171279, 2: 97104, 3: 96447, 4: 86806, 5: 20528, 6: 66330, 7: 34427, **8: 56689**}

{0: 3377, 1: 3224, 2: 23588, 3: 7646, 4: 21246, 5: 4127, 6: 13155, 7: 14260, **8: 49731**}

When I tried to find the no of people who are common to all three K-Mean Model (30, 45 and 60 features model), there are 33720 people. Therefore for a more targeted marketing, the company should spend more effort on targeting these 33720 people. For a more conservative approach, the company can instead assiduously target 41580 people, who are common to 30 and 45 features K-Means models.

MEAN VALUES OF FEATURES IN AZDIAS

Next, using 60 features K-Means model, I would find the mean values of the features for people in cluster 0 (also cluster 3 and cluster 6 since these clusters have the fewest customers).

It seems that for many KBA13 or KBA05 type features, the values for the three clusters are close to each other.

Following are the features that are best at distinguishing between customers and non-customers.

|  |  |  |  |
| --- | --- | --- | --- |
| **CLUSTER** | **0** | **3** | **6** |
| D19\_SOZIALES | Multi buyer 0-12 month | no transaction known | no transaction known |
| D19\_KONSUMTYP\_MAX | possibly versatile | inactive | inactive |
| D19\_KONSUMTYP | between gourmet and family | inactive | inactive |
| PLZ8\_ANTG4 | lower share of >10 family share | higher share | higher share |
| RT\_SCHNAEPPCHEN | higher share | lower share | lower share |
| D19\_VERSAND\_ONLINE\_DATUM | activity elder than 2.5 yr | no transaction known | no transaction known |
| D19\_GESAMT\_ANZ\_24 | between very low and low | no transaction known | no transaction known |
| MOBI\_RASTER | middle mobility | high mobility | high mobility |
| D19\_GESAMT\_ONLINE\_DATUM | activity elder than 2 yr | no transaction known | no transaction known |
| FINANZ\_SPARER | between high and very high | average | between average and low |
| D19\_VERSAND\_DATUM | activity elder than 2.5 yr | no transaction known | no transaction known |
| D19\_GESAMT\_ANZ\_12 | very low activity | no transaction known | no transaction known |
| D19\_GESAMT\_DATUM | slightly increased activity last 12 months | no transaction known | no transaction known |
| PLZ8\_ANTG3 | lower share of 4-10 family home | higher share | higher share |
| ALTER\_HH | 12-born in 1950-1954 older people | unknown | 17- born in 1975-1979 younger people |
| CAMEO\_HOUSEHOLD | comfortable household | between comfortable and less affluent | between comfortable and less affluent |
| BALLRAUM | 40-43 from urban center | 30-40 from urban center | 30-40 from urban center |
| GEBAEUDETYP | residential building without known household | mixed building | mixed building |
| D19\_BUCH\_CD | between double buyer and single buyer 0-12 month | between no transaction known and multi buyer 0-12 month | Multi buyer 0-12 month |
| D19\_LOTTO | Double buyer 0-12 month | no transaction known | no transaction known |

Feature means for cluster 0 (potential customers), 3 and 6 (non-customers)

There are no descriptions for D19\_SOZIALES and RT\_SCHNAEPPCHEN.

Customers are generally active consumers while non-customers are inactive. They also live in area with lower share of family home of >4 people, which mean they don’t generally live in shared housing. They also have elevated transaction activity for TOTAL POOL and for MAIL ORDER segment (makes sense!). They also have lower mobility and tend to be financial savers. They are generally born in 1950-1954 while non-customers are born in 1975-1979. They also live in more comfortable household (wealth status), live further away from urban center and in residential building while non-customers live in mixed building. They also purchase books, CDs and lottery products more frequently.

MEAN VALUES OF FEATURES IN MAILOUT TRAIN

For completeness of analysis, I would also tabulate the mean values of features for people in the MAILOUT TRAIN that my optimized model predicted to be customers and non-customers. For those predicted as customers, I only selected those predicted with probabilities of > 0.99 which amounted to 6586 people out of which 365 are correctly predicted to be customers. (365/532 = 68.6%). For those predicted to be non-customers, a probability of <= 0.04 is selected, amounting to 6898 people, of which 7 is wrongly predicted to be customers. Therefore, the number of potential customers and non-customers are pretty balanced.

I will start with non-KBA013 or KBA05 features first.

|  |  |  |
| --- | --- | --- |
| **Features** | **>0.99 (Potential customers)** | **<=0.04 (non-customers)** |
| D19\_KONSUMTYP\_MAX | possibly versatile | informed |
| D19\_KONSUMTYP | between versatile and gourmet | between informed and modern |
| PLZ8\_ANTG4 | lower share of >10 family house share | higher share |
| D19\_VERSAND\_DATUM | activity elder than 1.25 yr | activity elder than 1.6 yr |
| D19\_GESAMT\_DATUM | activity elder than 0.8 year | activity elder than 1.25 year |
| MOBI\_RASTER | middle mobility | high mobility |
| D19\_GESAMT\_ONLINE\_DATUM | activity elder than 1.5 yr | activity elder than 1.5 yr |
| FINANZ\_SPARER | very high | high |
| OST\_WEST\_KZ | west | split between west and east |
| PLZ8\_ANTG3 | lower share of 4-10 family home | higher share |
| CAMEO\_HOUSEHOLD | between prosperous and comfortable household | between comfortable and less affluent |
| BALLRAUM | 40-44 from urban center | 30-40 from urban center |
| D19\_BANKEN\_DIREKT | between no transaction known and multi buyer 0-12 month | Multi buyer 0-12 month |
| GEBAEUDETYP | residential building without known household | mixed building |
| D19\_BUCH\_CD | single buyer 0-12 month | double buyer 0-12 month |
| ANZ\_PERSONEN | more adult person in household | less adult person |
| D19\_LOTTO | Single buyer 0-12 month | between double buyer and single buyer |

Feature means for potential customers and non-customers in MAILOUT TRAIN

It can be seen that potential customers in MAILOUT TRAIN share common traits as potential customers (cluster 0) in AZDIAS. Here non-customers tend to be more active consumers, have slightly higher transactions in MAIL ORDER and TOTAL POOL, and higher financial saving habits, although still lower than that of potential customers. There are two other distinguishing features, namely, potential customers predominantly live in West Germany and there are more adult persons in their household.

For MAILOUT TRAIN, KBA013 and KBA05-related features have differentiated mean values between potential customers and non-customers.

|  |  |  |
| --- | --- | --- |
| **Features** | **Share magnitude** | **Share magnitude** |
| KBA13\_SITZE\_6 | higher share | lower share |
| KBA13\_SEG\_SPORTWAGEN | higher share | lower share |
| KBA13\_HALTER\_30 | lower share | higher share |
| KBA13\_ALTERHALTER\_30 | lower share | higher share |
| KBA13\_MERCEDES | higher share | lower share |
| KBA13\_HERST\_BMW\_BENZ | higher share | lower share |
| KBA13\_KMH\_250 | higher share | lower share |
| KBA13\_HALTER\_25 | lower share | higher share |
| KBA13\_GBZ | higher share | lower share |
| KBA13\_BAUMAX | lower share | higher share |
| KBA13\_ANTG1 | higher share | lower share |

Feature means for KBA05 and KBA13-related features for potential and non-customers in MAILOUT TRAIN

Potential customers tend to live in PLZ8 with higher share of > 5 seats cars, sports cars, Mercedes, BMW, and cars with max speed between 210 and 250 km/h, implying they are wealthier. Potential customers also live in PLZ8 with lower share of car owners aged between 21 and 31, implying they tend to be in older age bracket.

MAILOUT TRAIN

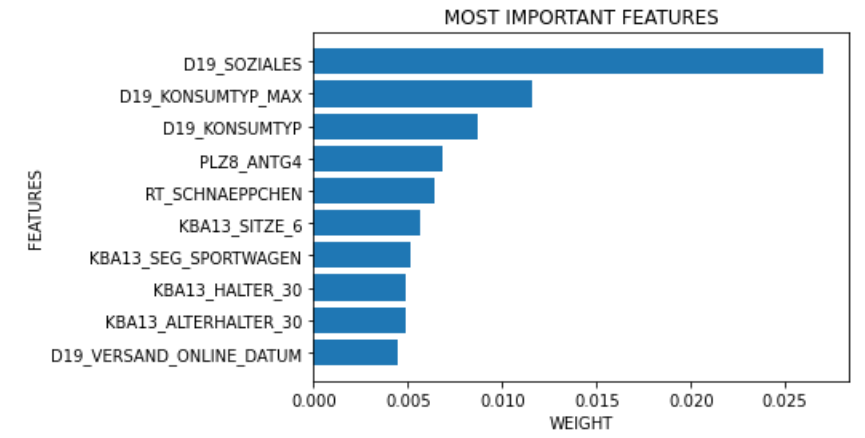
The mailout train data is an imbalanced dataset with only 532 positive responses and 42430 negative responses. I performed the same preprocessing steps earlier but I did not drop any row to prevent loss of information. I decided to use tree-based ensemble method because it will ignore redundant or correlated features in constructing ensemble of trees. While it is a greedy algorithm in that it chooses the best feature to split at that time but not the best overall, it is robust to overfitting and has proven to acquire the best score in Kaggle competition.

For hyperparameters optimization, I used Bayesian optimisation which is better than random search and grid search. I used stratified cross validation (since it is an imbalanced dataset) and split the mailout train into 5 folds. The mean AUC score over 5 folds is taken. I tried training Random Forest, Adaboost, Gradient Boost Classifier, Catboost and XGBoost and tested it on mailout test (Kaggle). Random Forest returned the worst result (0.5-0.6 AUC), Adaboost returned around 0.70-0.75, Gradient Booster and Catboost around 0.75-0.78. Lastly while XGBoost started out badly with 0.6 AUC, however Bayesian optimisation procedure was able to find hyperparameters combination that yield 0.79-0.803 AUC.

|  |  |
| --- | --- |
| **ENSEMBLE METHOD** | **ROC\_AUC SCORE** |
| Random Forest | 0.5 - 0.6 |
| Adaboost | 0.70 - 0.75 |
| Gradient Booster and catboost | 0.75 - 0.78 |
| XGBoost | 0.79 - 0.803 |

Range of ROC\_AUC score for tree-based ensemble method

Using the XGBoost model that yields the highest AUC (0.803), I proceeded to determine the most important features in predicting potential customers. Following are the 10 most important features with their corresponding weights.



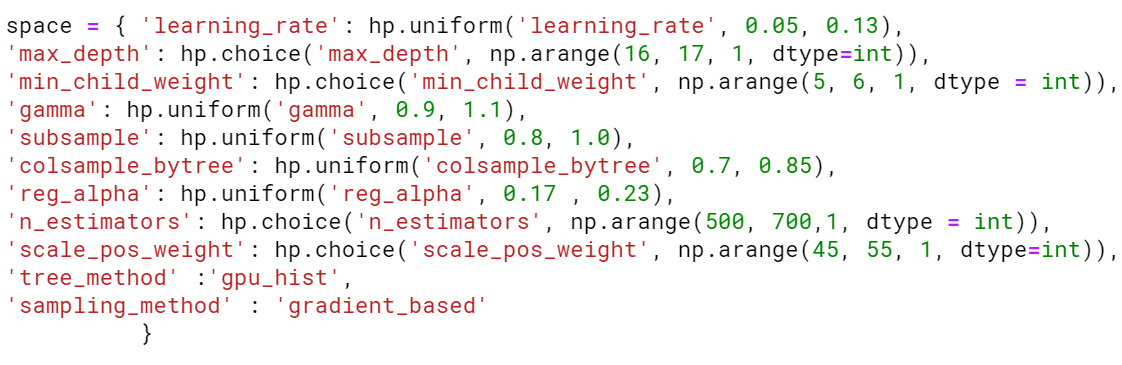
Top ten most important features as determined by XGBoost model trained on MAILOUT TRAIN

I realised that the imbalanced dataset with dearth of positive response examples is not a sufficiently good dataset to train my model on. While XGBoost has a hyperparameter called scale\_pos\_weight that can account for imbalanced dataset by increasing the weight of cost curve gradient updates of positive examples, yet this is still not sufficient. I thought of increasing the positive examples by using my best XGBoost model to predict potential customers in the AZDIAS and CUSTOMER dataset and adding those predicted to have positive responses (using >0.50 cutoff) to my MAILOUT TRAIN dataset. However after performing this procedure, my AUC score barely inched up.

OTHER APPROACHES

I then thought that maybe using the entire AZDIAS and CUSTOMER dataset is a better proposition. Since my 60-features K-Means analysis identified cluster 0 as the most likely customers in AZDIAS, I will label people in this cluster as positive (1) and the rest negative. As for CUSTOMER dataset, I will label everyone as positive (1). I utilised Bayesian optimisation again, but this time I will use MAILOUT TRAIN as my validation dataset. I arbitrarily chose the 130 most important features that were determined by my previous XGBoost model with AUC 0.803 as my features. This surprisingly has the effect of substantially raising my AUC score on MAILOUT TRAIN as compared to using all 355 features. I found that the max\_depth hyperparameter is the most important. Initially I got AUC of 0.82-0.84 by using max\_depth = 14. Subsequently, I realised that max\_depth 16 is the most optimal depth. As for other hyperparameters, I progressively narrowed my search range by looking up trained models with hyperparameter combinations that yield the best AUC score in both mailout train and mailout test. As such, I was able to arrive at a hyperparameters combination with max\_depth of 16 that produce the highest AUC score in the leaderboard, 0.88149. I have tried about 30 more times to best my result. However, my effort was futile and I persuaded myself to take satisfaction in my current result.

Following was my refined search space:

Refined search space for Bayesian hyperparameter tuning

Next, I found out that my best model predicted 22209 examples in MAILOUT TRAIN to be potential customers and able to predict 508 out of 532 examples correctly (using > 0.50 as cutoff). This may seem to be too many examples for it to be cost-efficient. Therefore I proceeded to find if we can find a higher cutoff that ensures many less examples are predicted to be positive yet still able to predict most of the 532 examples correctly. I found that by setting the cutoff at 0.99, I would predict 6586 positive examples and predict 365 out of 532 examples correctly. If I have access to the fixed, variable and marginal cost of marketing campaign, I would be able to calculate the cost-effectiveness more precisely. Devoid of such figures, I performed a back of the envelope calculation by dividing number of positive examples predicted by number of positive examples predicted correctly.

With 0.50 cutoff, the figure is 43.71 and with 0.99 cutoff the figure is 18.0. Lower figure is better. Or in other words, by reducing predicted examples by 70.35% , the positive examples I predicted correctly are reduced by 29.4%. Seems like a really worthwhile trade-off!