arvato
BERTELSMANN

Robert Kwapich

Customer Segmentation

report

Table of contents

Tables and figures

Tables and figures

DEFINITION

Project overview

The data and outline of this project was provided by Arvato Financial Solutions, a Bertelsmann subsidiary. The data itself is protected under terms and conditions in provided terms document, and cannot be shared or distributed publicly.

The main scope of the project is to analyze the demographics data for Arvato customers, a mail-order sales company in Germany. The goal is to enable customer-centric marketing campaigns through customer segmentation with the help of machine learning to pattern discovery.

First part of the project will help to understand the core customer base of the company in relation to the general population of Germany. It will provide necessary segment characteristics and descriptions pertaining to distinguished market sub-populations. It will provide an insight into which parts of the general population and their characteristics make a person more likely to be an existing customer.

Second part of the project utilizes results from the historical marketing campaigns, and aims at predicting the likelihood of a positive response from the marketing campaign. The positive response means becoming a client in some define future time. This part of the project is concerned with appropriate model building, optimization, validation and testing.

Provided data-sets have not been pre-cleaned and are accompanied by spreadsheet metadata files that describe various features collected, their codes, range of values and missing entries. All the supporting analyses and codes are hosted publicly on GitHub repository.

Input data

Provided data-sets have not been pre-cleaned and are accompanied by spreadsheet metadata files that describe various features collected, their codes, range of values and missing entries. There are four data files associated with this project:

`1. Udacity_AZDIAS_052018.csv`: Demographics data for the general population of Germany; 891 211 persons (rows) x 366 features (columns).

`2. Udacity_CUSTOMERS_052018.csv`: Demographics data for customers of a mail-order company; 191 652 persons (rows) x 369 features (columns).

`3. Udacity_MAILOUT_052018_TRAIN.csv`: Demographics data for individuals who were targets of a marketing campaign; 42 982 persons (rows) x 367 (columns).

`4. Udacity_MAILOUT_052018_TEST.csv`: Demographics data for individuals who were targets of a marketing campaign; 42 833 persons (rows) x 366 (columns).

Along with two spreadsheet files containing metadata files for the collected features:

1. Dias Information Levels - a top-level list of attributes and descriptions organized by informational category

2. Dias Attributes - a detailed mapping of data values for each feature in alphabetical order

All the supporting analyses and codes are hosted publicly on GitHub repository.

Evaluation metrics

The final evaluation metric is Area Under Curve (AUC) in the Receiver Operator Characteristics (ROC) for a response to the marketing campaign. The choice of metric is motivated by the fact that the data-set reflects significant class imbalance, and hence a simple accuracy measure would prove insufficient.

EDA and Pre-processing

Detailed descriptions and associated code is provided in accompanying GitHub repository.

Missing values

The first step in data cleaning (i.e. pre-processing steps) was to correctly identify the missing values in provided data-sets: `azdias`, general population of Germany, Arvato `customers` and `MAILOUT train/test` data-sets representing the results of marketing campaigns along with general features.

In particular all mentioned datasets have a mixture of values referring to lack of information, i.e. `NA` values. These values are not consistent and encoded through various annotations: `NA`, empty entries or codes like: `1,0` or `9`. Missing entries were standardized and encoded as `NA`. This allowed to estimate percentage of missing values per feature.

Undefined features

Subsequent step in data exploration was to identify the features defined in metadata files. Analysis revealed that the intersection of datasets and metadata information was not complete. In particular there were many features defined in provided datasets that didn't exactly correspond to the features defined in metadata. The opposite was also true: there were many features defined in meta but they were lacking entries in datasets.

Upon manual inspection it turned out that the feature names were often slightly changed, and hence exact matching was impossible. Filename `EDA/azdias_corrected_features.tsv` was created to provide a mapping between feature names defined in datasets and metadata files. This approach allowed to reduce by around half the number of seemingly undefined features. Finally, the remaining unidentified features were later inspected, and for the most part identified to be categorical.

Missingness filtering

Provided datasets were often incomplete and required curation. In order to prepare datasets for subsequent analysis steps, the degree of missingness was estimated, and filtering threshold have been applied to remove samples and features with excessive percentage of missing values.

For the first part of the project, datasets `AZDIAS` and `customers` were filtered with stringent thresholds of 30% per feature and 50% per sample (individual). This removed around 10% of features and around 25% of individuals from both datasets. The defined thresholds of 30% and 50% per feature and sample are arbitrary, and could be changed in confirmation analyses to identify whether they impact subsequent conclusions.

For the `MAILOUT` dataset, the established threshold were much less stringent due to the fact that this data has substantially less entries, and the response variables are vastly imbalanced,

and hence any potential data loss would have significant impact on the number of positive examples of response and the final AUC-ROC metric score. In addition, the final submission required all samples from `MAILOUT test` to be accompanied by relative likelihood of responding to a marketing campaign. For this reason the `test` datasets doesn't have per-sample filtering threshold, whereas for the `train` dataset the threshold is less stringent: 90% of missingness, which did not remove any samples. Feature-based missingness for `MAILOUT` dataset was set 30%, the same as `AZDIAS` and `customers` dataset, and removed six features.

Feature encoding

Features present in datasets were a mixture of numerical, ordinal, categorical and undefined types. First, all known feature types were manually annotated and placed in prepared `EDA/metadata_feature_types.tsv` file. Unknown feature types were assessed from the point of number of unique values, and the name of the unknown feature. The default data type for unknown feature is **categorical** type, as we cannot assume numerical properties, or a particular order between subsequent codes for the feature values.

Feature named `EINGEFUEGT_AM` was removed from all datasets as it had usually above 2000 unique values, and encoded some type of datetime variable. Since provided datasets contain other features encoding year, this seems a justifiable choice. Later features `ANZ_STATISTISCHE_HAUSHALTE`, `VERDICHTUNGSRAUM`, and `EINGEZOGENAM_HH_JAHR` were encoded as numerical, as they referred to unknown statistical property of the household (i.e. numerical value), some value of the volume, and the year which is a numerical value, i.e. can be regressed, and a fraction can have a meaning in real world.

The rest of the features were one-hot encoded to preserve the meaning, i.e. "interpretability" of the variable. An example of such a feature is `CAMEO_DEU_2015` which has defined 44 unique categories, whose interpretation cannot be numerical or ordinal. Categories defined by this example feature are referring to detailed classification of a particular person work/life orientation, and need to be interpreted in an "orthogonal" manner.

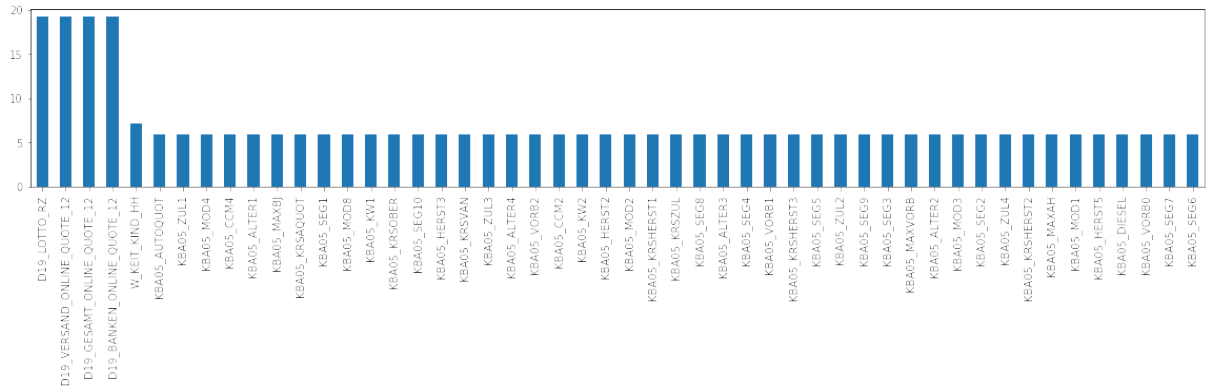


Figure 1: Missingness of features for combined AZDIAS and customers datasets after cleaning and pre-processing steps before imputation.

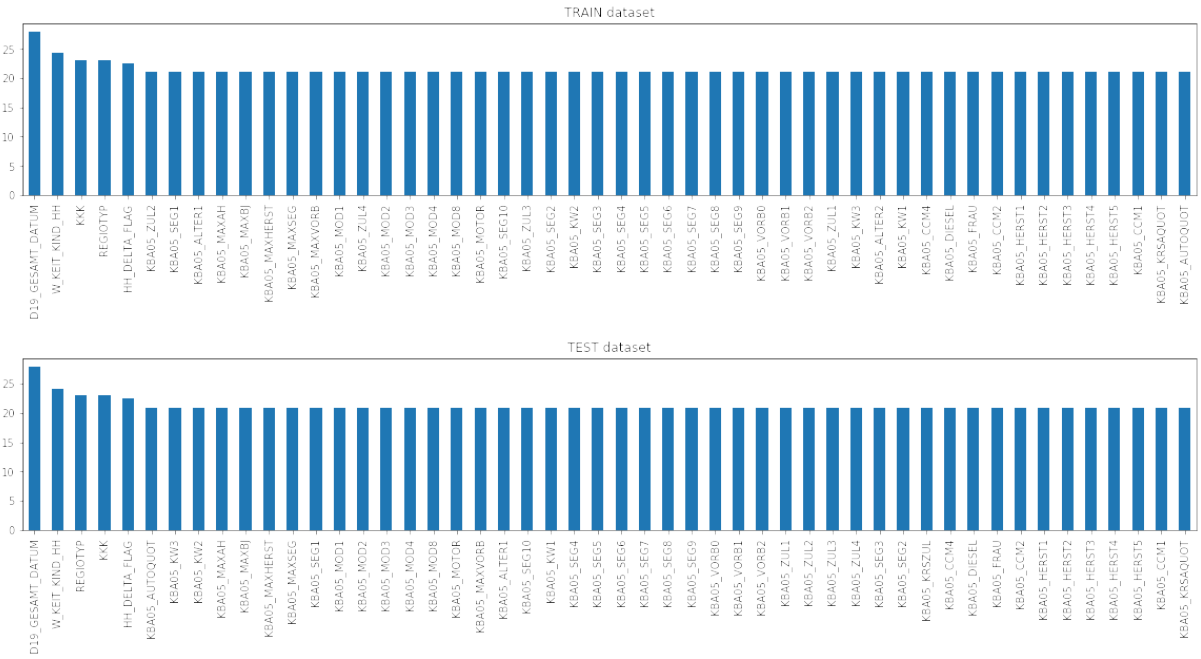


Figure 2: Missingness of features for MAILOUT datasets: test and train after cleaning and pre-processing steps before imputation.

One last step in feature encoding is feature encoding standardization. Namely, certain features use several labels to encode the same category, therefore lowering the specificity of the analysis, and introducing arbitrariness. An example of such a feature is `LP_STATUS_GROB` or `LP_FAMILIE_GROB` that collapse many initially defined labels into one category.

Concatenating data

Datasets `AZDIAS` and `customers` concerned with the first part of the project analysis, market segmentation, need to be concatenated (jointed) together. In order to achieve this the common set of featured needed to be establish. Overall, the vast majority of features between these two datasets agreed, except for the following features:

- Present only in `AZDIAS`: `KKK, REGIOTYP`

- Present only in `customers`: `CUSTOMER_GROUP, ONLINE_PURCHASE, PRODUCT_GROUP`

These features were removed from subsequent analysis.

Imputation

Parts of subsequent analyses required lack of missing data in provided datasets. This motivated the use of imputation methods - methods that could estimate the missing values utilizing the information present in provided dataset matrix.

The method used is a type of multivariate feature imputation, a more sophisticated **Bayesian** approach which models each feature with missing values as a function of other features, and uses that estimate for imputation. From the method documentation:

> *"It does so in an **iterated** round-robin fashion: at each step, a feature column is designated as output y and the other feature columns are treated as inputs X. A regressor is fit on (X, y) for known y. Then, the regressor is used to predict the missing values of y. This is done for each feature in an iterative fashion, and then is repeated for max_iter imputation rounds. The results of the final imputation round are returned."*

In practice the, due to the vast number of features, the parameter n_nearest_features, that controls the number of other features used to estimate the missing values of each feature column, was reduced to 50-100 in order to make it more computationally tractable. This allowed to impute missing features focusing only on the most correlated (i.e. relevant) portion of the features, and significantly speed-up the computation. Still, the imputation part, beside hyper-parameter tuning, one of the more computationally intensive steps in the analysis.

Part 1: Customer Segmentation

Overview

Customer segmentation is an unsupervised learning problem. Namely, the objective is to identify potential patterns, and stratify obtained data into clusters based on (dis)similarities of subjects in the dataset.

Methodology

Feature scaling

The previous step ("EDA and Pre-processing") prepared dataset by curating its entries (encoding, standardizing), and imputing missing values. Before applying below method, it is necessary to standardize each feature to put equal importance in terms of variance explained. Otherwise features that encode numerical values, like year, and which have values 19xx-20xx, would overwhelm other features in the unit scale.

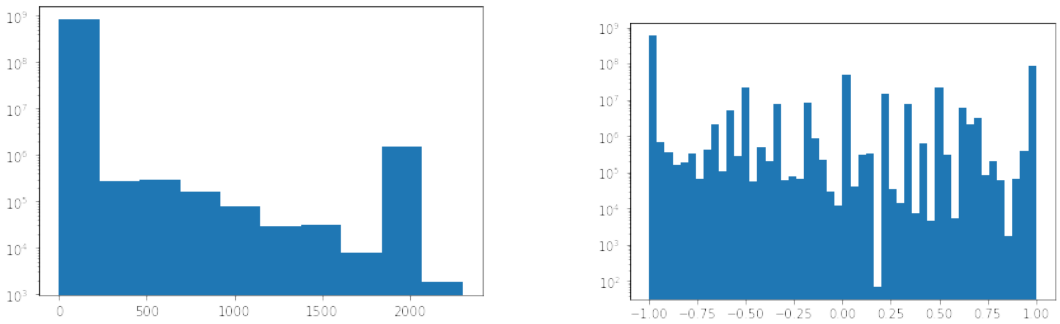


Figure 3: Histograms of joint and curated AZDIAS+customers dataset before (left) and after (right) scaling. The Y axis is in log scale and shows the counts of values falling into 50 bins corresponding to feature values ranges on X axis. The range of possible values for features has been restricted with min-max scaling to (-1, +1) range.

Correlation analysis

Before employing dimensionality reduction methods in subsequent steps, it is beneficial to perform a basic correlation analysis to identify correlated features. Due to the fact that the curated dataset has many features, especially after one-hot encoding categorical variables, it is necessary to identify and remove highly correlated ones. Subsequently features that correlate more than 0.9 of Pearson Correlation are identified and one out of pair of features is removed.

Figure 4 shows the plot of correlation matrix for AZDIAS+customers dataset. It is clear that not many features are either positively or negatively correlated. This resulted in removal of total 16 redundant features shown in Table 1.

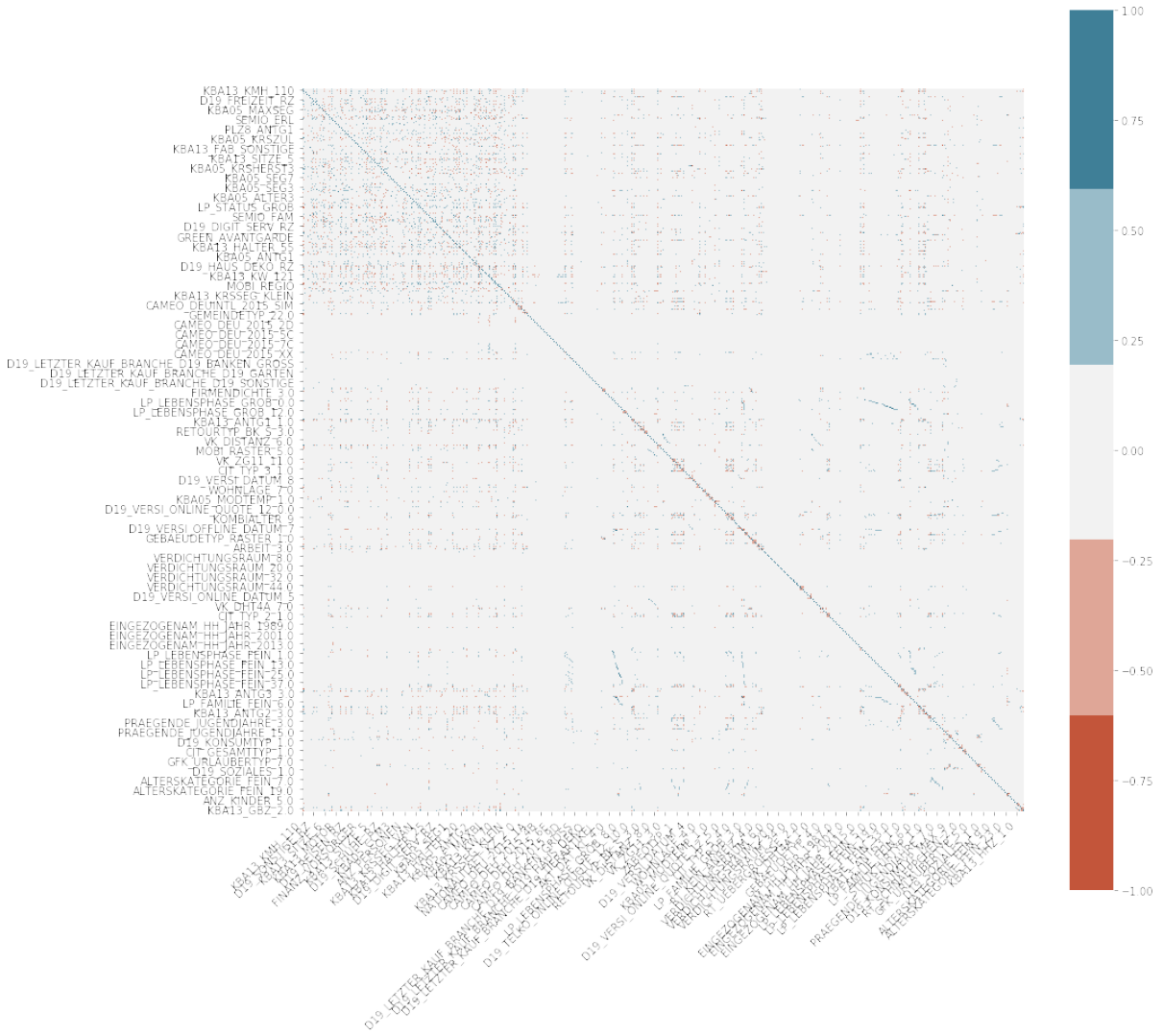


Figure 4: Correlation matrix of features in joint AZDIAS+customers dataset. Many present features exhibit negligible or small correlations.

This allowed us to quickly to assess the features that convey similar portion of information and remove all but one from a “bag” of correlated features.

Table 1: Features identified to be correlated above 0.9 value of Pearson correlation. Left column identifies features kept in the datasets, right column identifies a single feature or multiple features that were removed from subsequent analyses. For the purpose of this analysis these features could be treated as equal, and hence interchangeable.

Feature Kept	Removed features
KBA13_FAB_SONSTIGE	KBA13_HERST_SONST
KBA13_ALTERHALTER_61	KBA13_HALTER_66
D19_GESAMT_ONLINE_QUOTE_12	D19_VERSAND_ONLINE_QUOTE_12
KBA13_KMH_250	KBA13_KMH_211
ANZ_STATISTISCHE_HAUSHALTE	ANZ_HAUSHALTE_AKTIV
CAMEO_DEUINTL_2015	CAMEO_DEUINTL_2015_SIM CAMEO_DEUG_2015,
D19_TELKO_ONLINE_QUOTE_12_0.0	D19_VERSI_ONLINE_QUOTE_12_0.0
CJT_TYP_3_5.0	D19_KONSUMTYP_MAX_9 CJT_TYP_4_5.0
FIRMENDICHTE_5.0	GEBAEUDETYP_RASTER_5.0
LP_LEBENSPHASE_GROB_0.0	LP_LEBENSPHASE_FEIN_0.0
LP_FAMILIE_GROB_2.0	LP_FAMILIE_FEIN_2.0'
LP_FAMILIE_GROB_1.0	LP_FAMILIE_FEIN_1.0
KBA05_MODTEMP_5.0	KBA05_HERSTTEMP_5.0
D19_KONSUMTYP_MAX_8	D19_KONSUMTYP_9.0

Dimensionality reduction

To further facilitate the customer segmentation through clustering techniques, the input dataset could be reduced in the number of features. I've identified obviously correlated features, but one additional thing is to express features as a linear combination of other features. It would be benefitial to order them by some importance metric. This is exactly what Principal Component Analysis is performing: creating features expressed as ortogonal (independent) entities described by linear combination of original features, and orders them by the amount of variance explained.

If a feature is constant for all samples, then it doesn't tells us anything about the variability and potential clusters. And hence it would explain the least amount of variability. Analogically if there there are variables that jointly describe some physical entity relating to customer segments, they would jointly explain an independent feature which would need to be interpreted.

In my analysis I decided to re-express the AZDIAS+customers with the number of principal components that explain 80% of the original data variability. Since the analysis is not aiming at stratifying customer sub-segments, and only providing a broad overview, this is a justifiable choice. With this goal, the original dataset with 892 features (many of which were created by categorical one-hot encoding), was reduced to 215 most important features that are "mixture" (linear combination) of original features.

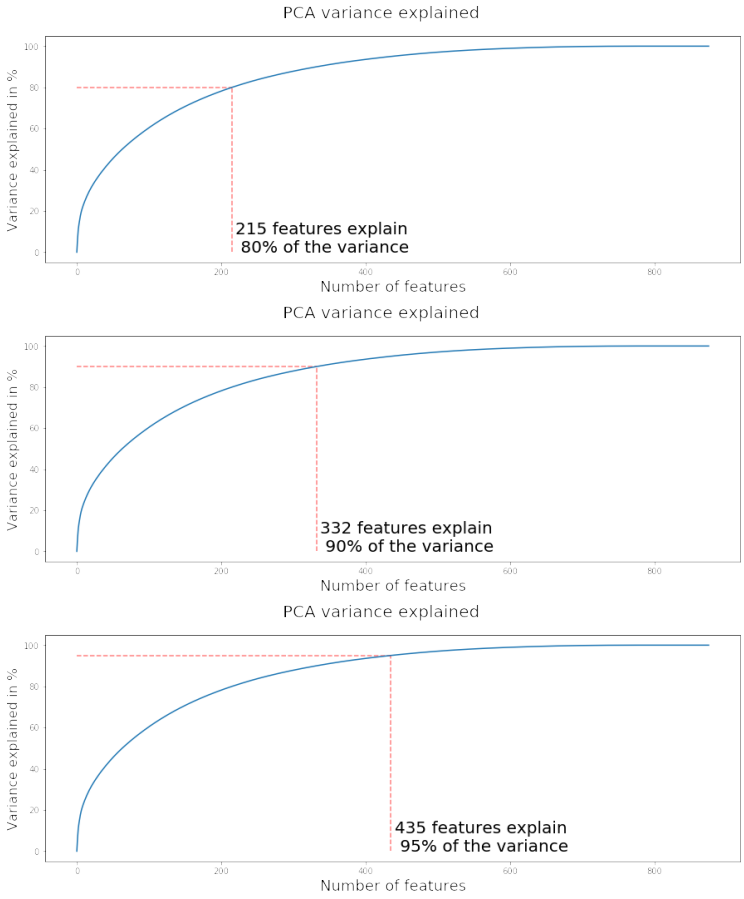


Figure 5. Variance explained by different number of features in AZDIAS+customers dataset.

Clustering

In order to find the customer segments, the curated dataset of Germany general population and Arvato customer base needs to be clustered. For the project I decided to perform K-mean clustering.

First, I employed the "elbow method" heuristics for finding the optimal number of clusters. That is I performed k-means clustering for number of clusters in the range 2-25, and computer metrics, such as: Calinski-Harabasz Scores and Sum of Square Distances. The main objective of the heuristics is to find a point (number of clusters) that minimize these metric values, but with an observation that this value will reach 0 in number of clusters is equal to the number of number of data points in the dataset. Hence, the "elbow", a point where the metric speed of decrease (i.e. derivative) reaches plateau, i.e. for increasing the number of clusters further we don't obtain significant metric reduction.

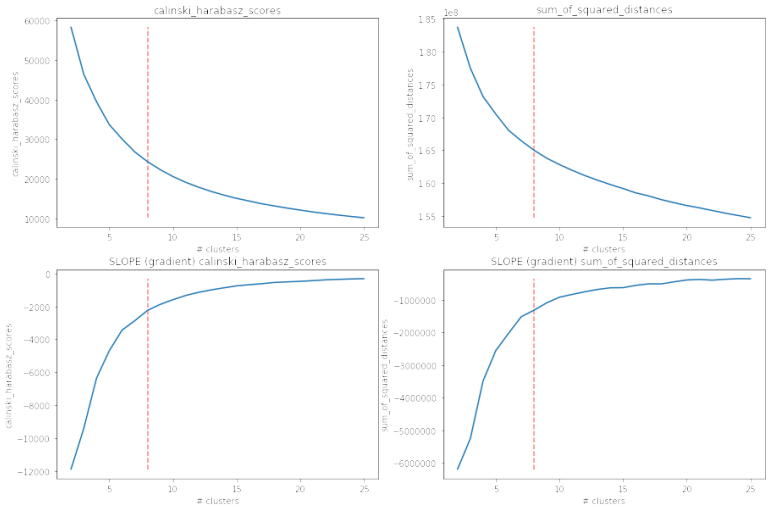


Figure 6. Calinski-Harabasz (first column) and Sum of Square Distances (second columns) metrics displayed on Y axis for different clusters in range 2-25 on X-axis. The second row show the slope (derivatived), i.e. the speed of decrease of a particular metric. Vertical red bar has its origin in x=8, the number of clusters chosen to segment the customer base.

The number of clusters from Figure 6 lays in the range 6-10, and hence 8 seems like an optimal solution. Further, for this number of clusters a **cluster stability test** was performed. Namely, for random samplings of our dataset we want to compute a score that reflects how close this clustering is consistent, and how often it changes depending on the random sampling. The metric used was the Rand Index for cluster assignment . Its definition could be found on scikitlearn page, 0 reflects random cluster assignment , whereas 1.0 reflects exact clustering each time. For `arvato+customers` dataset the cluster assignment reached the Rand Score of 0.98.

Results

Cluster relative proportions

The meaning of the clusters is going to be defined in the next sub-section. Figure on the right shows the relative proportions of clusters of `azdias` (general population of Germany) and `customers` the customer base of Arvato.

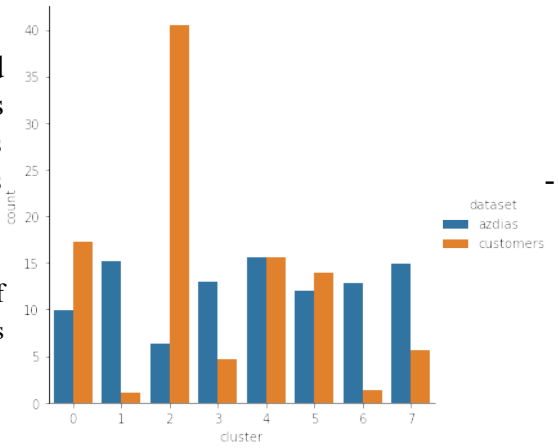


Figure 7. Relative proportions (summing to 100%) of azdias (general population of Germany) and customers (the customer base of Arvato) datasets.

We can observe the following:

Clusters 1, 3, 6, 7:

- these are the segments of general population for which Arvato has relatively less customers. Meaning this segment of population is **less likely to use the services**, as customers from this segment are less prevalent in our customer dataset.

Clusters 0, 2:

- these are the segments of general population that are over-represented in Arvato customer base. Meaning, **this segment of population is much more likely** (constitutes higher percentage) to use Arvato services.

Cluster 4,5:

- this is the segment of population that is equally represented in customers dataset with relatively high percentage. Meaning this is also an important from the business perspective segment of the market.

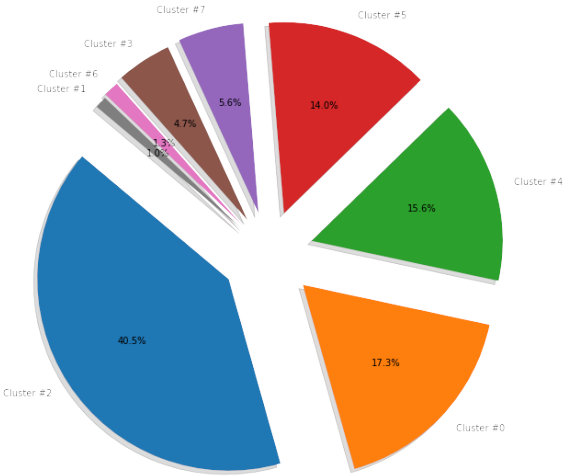


Figure 8. Relative proportions of identified clusters in Arvato customers dataset. Half of the distinguished clusters (2,0,4,5) represent over 85% of the overall Arvato customer base.

Principal Component Feature meaning

It is possible to decompose each principal component and identify the top feature contributors defining the axis. Hence, it is possible to interpret the dimensions defining the clusters. Detailed and technical discussion is placed in accompanying GitHub repository. Here I only discuss and show the major defining features of each principal component. I limit the discussion to top 5 Principal Components.

PCA #1

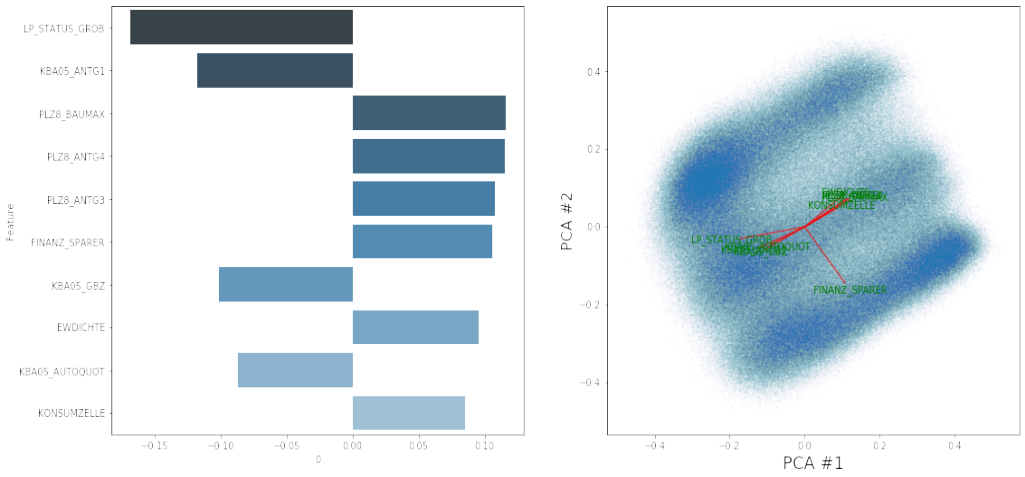


Figure 9. PCA #1. Left: decomposition of top 10 feature contributing towards understanding this feature. Right: PCA plot for selected axes, annotated with arrows pointing in the direction of a particular feature change.

PCA #1 corresponds roughly to a **general wealth of an individual**. Positive high values of this PC axes correspond to wealthy individuals that are not savers, have multiple cars, live in wealthy neighborhoods or cities of high population. Contrary, negative values correspond to poor individuals, that are strong savers, don't have a lot of cars, might live in less densely populated areas, like countryside.

PCA #2

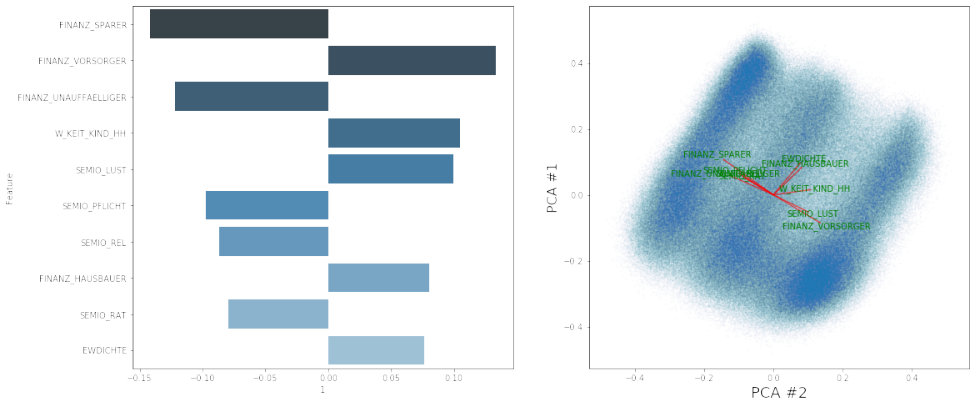


Figure 10. PCA #2. Left: decomposition of top 10 feature contributing towards understanding this

feature. Right: PCA plot for selected axes, annotated with arrows pointing in the direction of a particular feature change.

PCA #2 corresponds roughly to **general money management** skills and orientation (saving/ spending). High values of this PC axes mean households in high population density locations,most likely without children, with people oriented on money saving, with "remarkable" money management schemes, that are rational, and most likely traditionally religious. Conversely low values of this PC axes means people within households that most likely live outside of high density locations, with children, not particularly focused on money savings, but concerned about "financial preparedness" (so wise spending), probably more swayed by emotions and less traditionally religious.

PCA #3

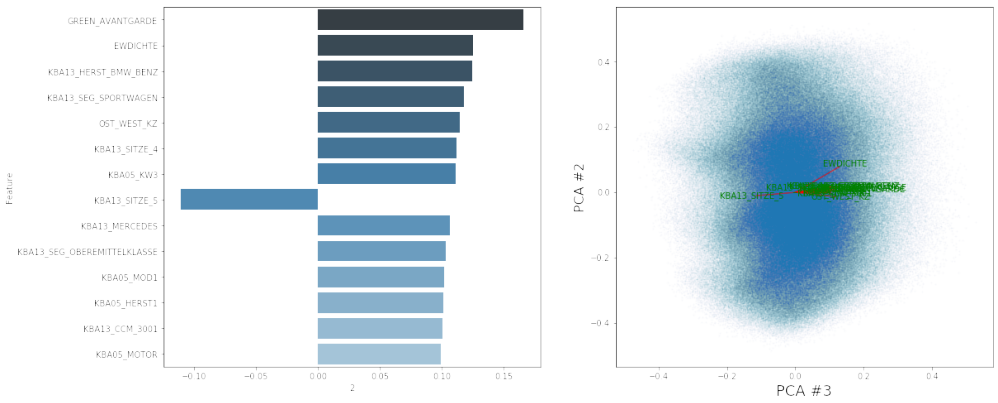


Figure 11. PCA #3. Left: decomposition of top 10 feature contributing towards understanding this feature. Right: PCA plot for selected axes, annotated with arrows pointing in the direction of a particular feature change.

PCA #3 roughly corresponds to **affinity to luxurious products**. High value of this PC axis correspond to customers within household that consider themselves (or have been classified) as "Green avantgarde", most likely coming from western germany, They drive very expensive cars, with number of seats most likely below 5, and high prevalence of sports cars, or cars from top manufacturers. Whereas low values of of this PC axis mean the oppositive of high PC values: conversely people that most likely leave in eastern germany, do not consider themselves "green", have inexpensive cars, usually family size that can accomodate more than or equal of 5 people (5 seats).

PCA#4

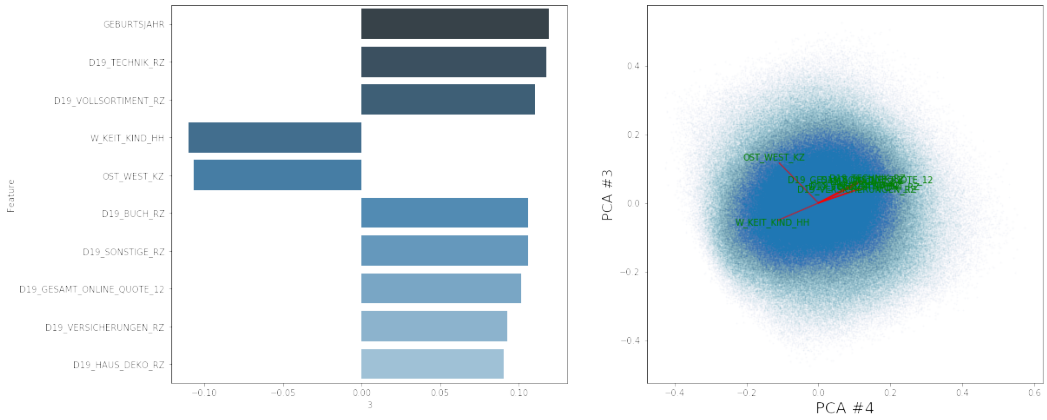


Figure 12. PCA #4. Left: decomposition of top 10 feature contributing towards understanding this feature. Right: PCA plot for selected axes, annotated with arrows pointing in the direction of a particular feature change.

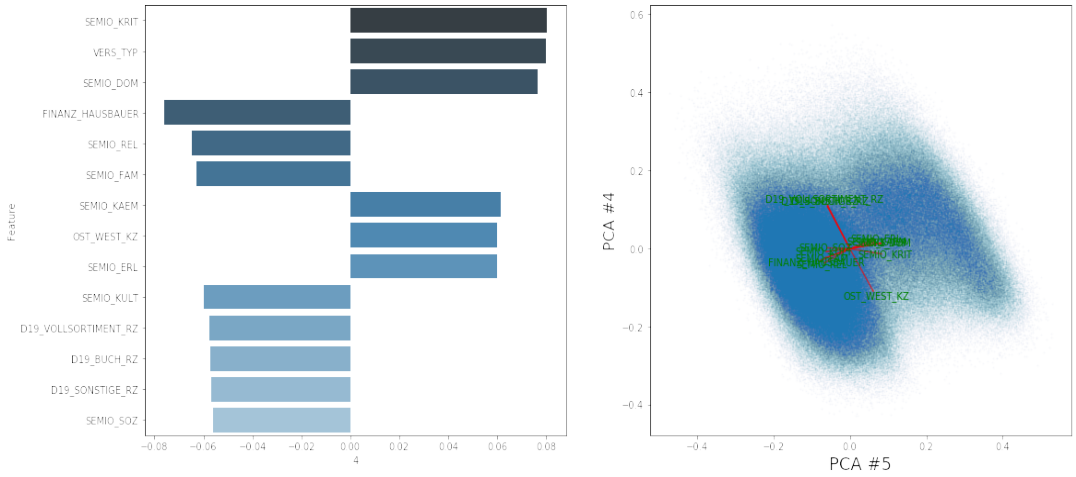
PCA #4 roughly corresponds to **age or generation** of a customer. With high values of this PC axis meaning young people, with activity predominantly over the web, rather than through mail-order. This group is likely to have children.

This group is less likely to have expensive cars, likely smaller `KLEINWAGEN`, and is also most associated with eastern parts of Germany, where, due to historical reasons, the accumulated wealth is smaller, than western part. This group is more likely to live in a zip code are (PLZ8) that comprises 6-10, and >10 family houses. Conversely low vlaues of this PC axis means older people, more likely to use mail-order placement than online purchases. Also they would likely be more likely to have more expensive cars, and be more prevalent in western Germany. They are much less likely to have children (small children) in the house.

PCA #5

PCA #5 roughly corresponds to the MAIL-ORDER purchase likelihood (associated with social orientation) versus the critical, fightful, dominant online shoppers. High values of this PC axis represent a segment of population oriented towards culture, familiarity, religion, owning a house, whose major attitudes don't include aggressive/fightful attitude, low critical thinking. However they are willing to accept risks, but they are also socially (community) minded. They are a multibuyers of COMPLETE MAIL-ORDER OFFERS. Conversely, low values of this P axis represent a segment of population that includes more self-oriented individuals, less likely to COMPLETE MAIL-ORDER OFFERS, less: religious, community-driven, likely from WEST germany with high criticial, dominant and fightfull atitudes.

Figure 13. PCA #5. Left: decomposition of top 10 feature contributing towards understanding this feature. Right: PCA plot for selected axes, annotated with arrows pointing in the direction of a particular feature change.



Cluster composition

In order to interpret the clusters, it is first vital to understand what principal coordinates, i.e. the linear combination of original data features, are most prominently defining these cluster. Thankfully it is quite straightforward with PCA and k-means clustering.

In the previous section I've decomposed each PC 1-5 into at least ten original features, and provided interpretations to the meaning of this feature. Now it is time to use the centroids from k-means, each described in PC dimensions. Figures 14 and 15 show heat-maps of PC axes values vs. identified clusters. Through these plots it is easy to distinguish what primary features and its values (positive or negative) primarily impact a particular cluster. Looking at figure 14 we see that to the most extent only top 5 features (shown in figure 15) mainly define these clusters.

Below is a brief description of each cluster in terms of PC components. In the next section I provide interpretations to these clusters. **Note:** PCs are numbered from 1-169. Clusters are numbered from 0-5.

1. **Cluster_0**: is described hugely by strongly negative values of PC1, and to a some degree **PC3**, while being anti-correlated with positive PC2 values that has mediocre impact on this cluster.

2. **Cluster_1** is described primarily by positive PC1 values and negative PC2/PC4

3. **Cluster_2** is described primarily by positive P2/P3 values and negative PC1 values

4. **Cluster_3** is described primarily by negative PC2 and PC1 with strongly negative PC2 values

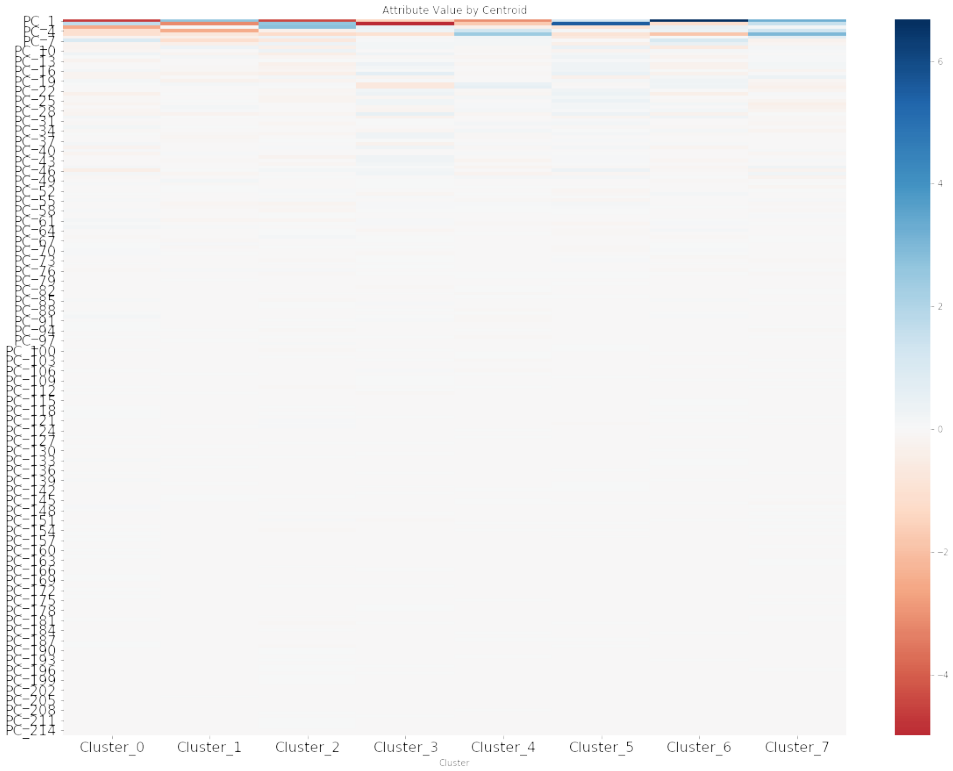


Figure 14. Heatmap showing decomposition of PC axes (Y axis) and its values represented by color (red for negative, white-gray for close to zero, and blue for positive) among identified clusters (X axis). Wee see that for all 215 PC dimensions, only the top 5-10 contain the most amount of information.

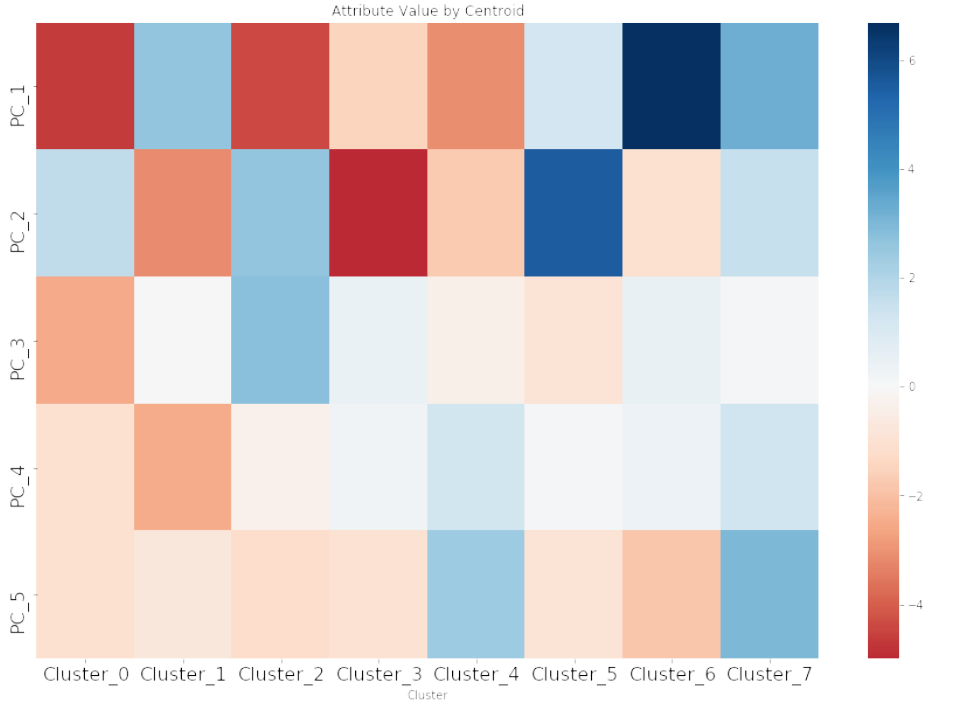


Figure 15. Heatmap showing decomposition of top 5 PC axes (Y axis) and its values represented by color (red for negative, white-gray for close to zero, and blue for positive) among identified clusters (X axis). Wee see that for all 215 PC dimensions, only the top 5-10 contain the most amount of information.

5. **Cluster_4** is described primarily by negative PC2 and PC1, but comparing to cluster 3, less strong values for PC2, slightly more importance in PC1

6. **Cluster_5** is described primarily by positive PC2 and PC1 values, with strong PC2 values

7. **Cluster_6** is described primarily by positive PC2 and PC1 values, with strong PC1 values

8. **Cluster_7** is described primarily by positive values in all considered PCs (1-5), without any strong values

Customer segments and their importance

Combining the information from original data features, principal coordinates and cluster composition it is possible to identify and describe the segments (i.e. clusters) of the customers and general population of Germany.

First, I provide a brief characteristics of each identified cluster:

1. **Cluster_0**

A population that is rather poor, strong money savers, without numerous cars (or any expensive ones), and lives in less densely to medium populated areas. They are most likely traditional and and religious. In higher likelihood to live in east Germany. Do not consider themselves "green avant-garde"

2. **Cluster_1**

A population of rather wealthy individuals that are not oriented on saving, but on "financial preparedness" (wise spending?). They can have multiple cars, likely to live in wealthy neighborhoods. They might have children, often less traditionally religious, more swayed by emotions. More likely to be slightly older people (not like young and poor parents, but rather wealthy mid 30s parents).

3. **Cluster_2**

A population of not rich individuals (less likely to have MULTIPLE luxury cars). They are Strong savers with money management skills, rational, most likely religious. Similar to cluster 0, but more wealthy, live in less populated areas and might come from Western Germany. Also with an affinity towards luxury cars (but do not have them a lot, but an affinity towards luxury).

4. **Cluster_3**

A population of rather not wealthy individuals, most likely with children, not particularly focused on money savings, considerate about financial preparedness

5. **Cluster_4**

Comparable to population from **cluster 3**, but the proportions are shifted. They are usually less wealthy, more focused on money saving. They are also more swayed (influenced) by emotions, but consider themselves wise spenders.

6. **Cluster_5**

Corresponds to population most likely without children with strong orientation on money saving and on money management. Rational and traditionally religious and less likely to perform purchases through mail orders, rather online.

7. **Cluster_6**

A population of very wealthy (**the most wealthy** customer segment), generally comprising of older people, but they are also less likely to use mail-order online placement in place for online purchases.

8. **Cluster_7**

A population oriented toward family and likely to have children. They are usually on the younger side of life, moderately wealthy with moderate to small cars. There is small positive correlation (association) to being in Eastern Germany. They are more likely to use mail-order offers, but are also buying online. Live in rather high density areas, and are somewhat oriented towards money saving. Usually rational with traditionally religious beliefs.

Part 2: Predicting marketing campaign success

Overview

This part of the projects is concerned with modeling customer responses to marketing targeted marketing campaign. In particular it is concerned with building a model that takes the population data, and outputs relative likelihood of a person to respond positively to a marketing campaign, i.e. of becoming a customer. This section is only going to briefly discuss models and the scores. For detailed technical comments and code please see the associated GitHub repository.

Methodology

Feature scaling

Feature scaling is analogical to the feature scaling described in Part 1 of this report. The only difference is that our input dataset is different: `MAILOUT train/test`. These datasets contain one additional field: `RESPONSE`, which doesn't exist (purposefully) for test dataset. Features were scaled with min/max scaler in the range -1 to +1.

Benchmark

In order to compare several models tested for this part of the project, the benchmark metric was established as ROC-AUC: Receive Operator Characteristic: Area Under Curve. The main reason for this choice of metric is that the RESPONSE variable exhibits huge imbalance (see figure 16). Namely, we have disproportionately more examples of negative cases, where a customer didn't respond to a marketing campaign, than positive ones. A simple accuracy measure would prove inefficient for model building and subsequent tuning.

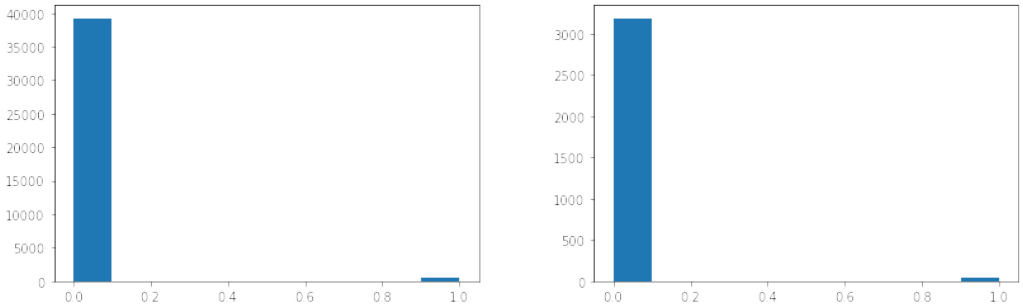


Figure 16. Imbalance of positive (1) and negative (0) cases among train dataset, that was further split 92.5% train (left panel) and 7.5% validation (right panel).

Results

Detailed technical comments and code are hosted in the associated GitHub repository. Instead here I only discuss the resulting AUC scores, the way the models were tuned and implication of the results.

Model 1: linear learner

Two linear models for binary classification with objective to maximize precision at the target recall were trained. The first and the second models were identical to almost all parameter, except the latter introduced a balanced weight matrix to account for class imbalances. The first model scored 0.67178 AUC, the second 0.62746 on the test set in the Kaggle competition. These models established a baseline AUC for more sophisticated models used in subsequent part of the analysis.

Model 2: Gradient boosted trees (XGBoost)

XGBoost library of gradient boosted trees was used to train a tree-based classifier. In fact this step involved a **hyperparameter** tuning jobs that spawned 100 models, and have chosen the one that performed best in terms of AUC score on a separate validation score.

The model reached a score of 0.78399 in the Kaggle competition. Comparing this score to the Kaggle leader-board it is still clear that model could be improved, how not by much. The top performing submission scored 0.81063 AUC, which is not a significant improvement. Overall archieved score of 0.78399 is significant improvement over random guess (0.5), but still lacking distinct certainty in prediction, as AUC of 0.9 and higher would give.

With XGBoost tree-based classifier it is possible in addition to **identify the most significant features** deciding on the outcome of the classification. From Figure 17 we see that the most important features refer to social status, vacation habits, consumption and shopping types.

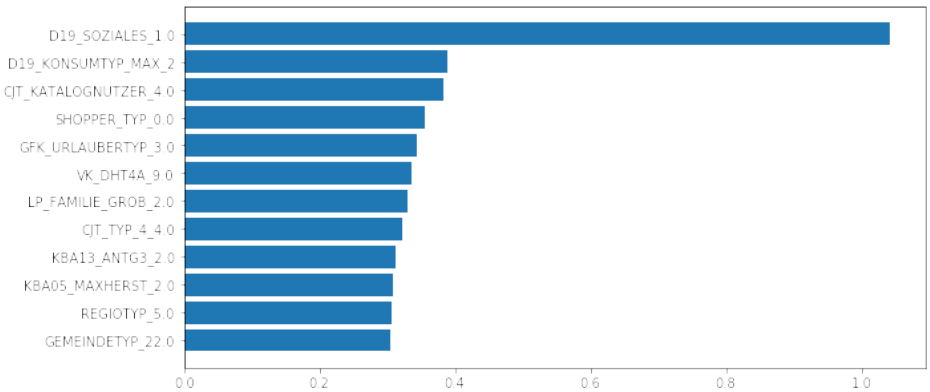


Figure 17. Top twelve features that exhibit highest impact on RESPONSE variable, and hence have highest impact on the decision of positively responding to the marketing campaign.

Comments

In theory mode models could have been tested, like very flexible densely-connected neural networks or random forest classifiers. In practice however, the XGBoost methods, due to the

boosting nature of the method, are considered state of the art methods of classification, and especially in the case of significant class imbalance. Due to limited time resources proposed solution achieves very acceptable AUC score, unlikely to be outperformed by other methods.

In particular it is worth re-iterating that the pre-processing steps also might have an impact on the overall accuracy. If all missing entries were to be imputed, instead of defining thresholds above which a particular feature is removed, then the resulting AUC score might change. Also, since XGB boosted-trees, unlike linear models, can work with missing values, it might be worthwhile not to attempt imputing the dataset, or remove missing values, but rather work with this imperfect dataset and then assess the AUC score. More extensive optimizations, that would take additional tens of hours would likely improve the overall AUC score closer to 0.8. However currently existing model is providing very close AUC with time-saving optimizations.

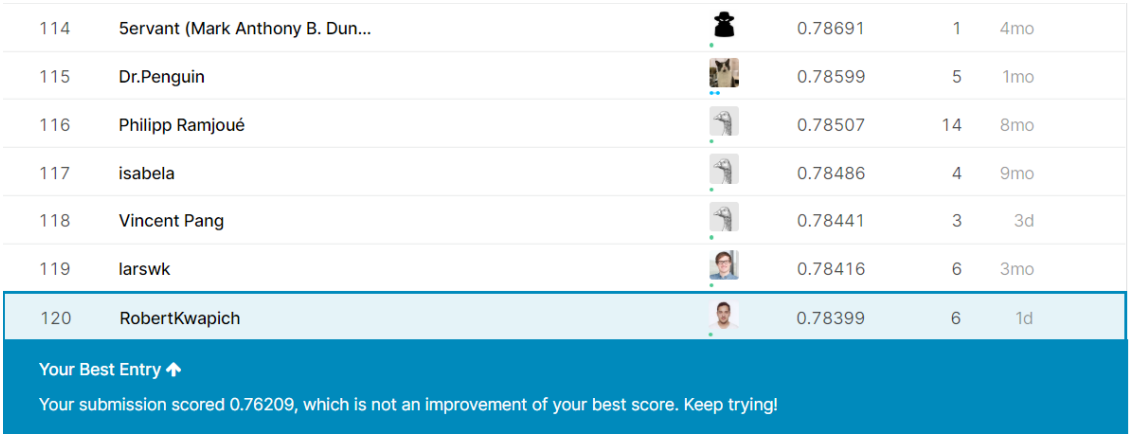


Figure 18. Final AUC score of 0.78399 for supervised problem of predicting customer response for marketing campaign.