Problem Overview

Customer acquisition is a key aspect of business growth; to be efficient, a firm must target its marketing efforts to true potential customers, rather than advertise to the general population. But how can a business discern which individuals from the general population are more likely to become its customers? This project addresses this question, and whether the customer acquisition strategy of a Mail-Order company in Germany can be made more efficient using Machine Learning (ML) techniques. To this end, a vast amount of demographic data has been made available for the general population of Germany, as well as from the firm's customers. Additionally, the results of a recent mailout campaign have been provided. All in all, the givens for this project are three datasets contained in six files as follows:

- 1. ***Udacity_AZDIAS_052018***.csv: Demographics data for the *General Population* of Germany; 891 221 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; 42 962 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).

Two ancillary spreadsheets: ***DIAS Attributes - Values 2017.xlsx***, and ***DIAS Information Levels - Attributes 2017.xlsx*** provide descriptions for most, but not all features in the data.

Problem Statement: The *General Population* and *Customers* files will be used to establish what makes customers stand out from the general population. Data visualizations and unsupervised learning ML techniques will be used to discriminate customers from non-customers. Then, the *Mailout* files will be used to train a predictor, and to test its predictions in a Kaggle Competition.

The essence of this project is didactic in nature; it is the capstone project culminating a series of assignments that aims to provide a well rounded understanding of ML techniques for someone interested in Data Science. First, we'll explore and preprocess the general population and customer datasets. We'll handle problems with the data such as missing entries, encoding and outliers in order to implement meaningful comparisons. We'll implement visualizations to gain insight as to how the features' distributions of the general population and the customers differ. Second, we'll carry out unsupervised learning ML techniques, namely, Customer Segmentation using Principal Component Analysis and K-Means Clustering to examine what they reveal about the characteristics of the customers. Third, we'll fit a variety of binary classifier models to the mailout campaign data, and fine-tune the parameters of one of these models as a supervised learning ML approach. Finally, we'll test the goodness of the tuned fit model by evaluating how well it predicts which individuals are indeed customers. Our goal is to conceive a targeted advertising campaign based on recommendations from the data itself, as exposed through ML techniques, in order to achieve a higher hit-ratio in marketing than is otherwise possible.

The implementation relies on Scikit-Learn algorithms and models, with some Python 3.0 code snippets and libraries. The project was completed on an Apple Computer iMac 24-inch with an M1 chip (8-core CPU, 8-core GPU) running macOS Version 12.0.1 (Monterey). Data wrangling, modeling took place on a Jupyter notebook of the Anaconda Navigator 2.0.4 suite of utilities. Editing and debugging with Atom 1.5.8. Code and notebook are available at GitHub Repository.

Data Disclaimer: Due to the sensitive and proprietary nature of the demographics data used in the project, the data will not be provided. The data was considered exclusive to complete the project and not for any other purpose, as according to the Terms & Conditions associated with this project on Udacity.

Data Exploration

Missing Data: The first challenge we faced was the amount of missing data, and the surprisingly structured character of the missing data as a series of jumps — Fig. 1.

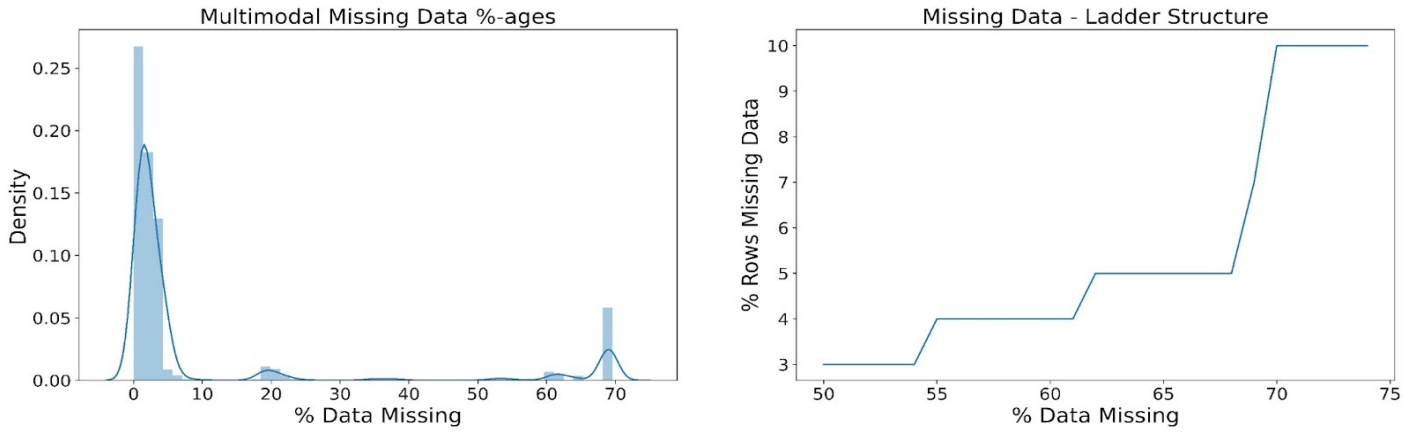

Figure 1 — Ladder Structure of Missing Data.

Figure 1 shows 10% of the general population rows missing 70% or more of their entries, 5% missing 63% or more, and so on. The customers' data is similarly affected. We don't know if the missing data is *Missing Completely At Random* (MCAR), *Missing At Random* (MAR), or *Missing Not At Random* (MNAR). This distinction is instrumental in deciding whether or not to apply listwise deletions, i.e., discarding rows.

The jumps may result from mixed data collection efforts, e.g., multiple surveys with separate questionnaires, or surveys across time, with later ones introducing a different set of questions. Joining such disparate observations into one big data file would naturally result in missing data with the structure depicted in Figure 1. This may also be surmised from the fact that most features are categorical and include heavily populated slots for ***'unknown'*** values, thus exposing the missingness of the blank entries as MCAR in type. Quintero et al (2018) find that *"A good example of data MCAR occurs when some subjects of study neglected to answer a question in a survey because they did not see the question" (1.1).* Here we are faced with a trade-off between discarding rows missing too many entries, or filling the data with educated guesses by imputing or regressing the features. To deal with the missing entries, Lokesh (2021) argues that *"When the cause of missing data isn't random, listwise deletion is another source of concern"* (*Listwise Deletion*). Hence, listwise deletion, or the discarding of rows, is a viable strategy for us here, since the missing values are not due to participant characteristics (demographics), nor feature characteristics, but arise completely at random.

Given our suspicions that the data origins are mixed, it is not wise to discard any data at all. There may be a small subset of columns resulting from a more focused survey that contains the most pertinent customer data, with excessive blanks elsewhere across the rest of the features. Lest we end up discarding the most meaningful piece of the data, it makes sense to solve our problem through imputation. This is the chosen route; no rows were discarded.

Preprocessing: The datasets consist mainly of ordinal pre-encoded features with labels starting at 0 or 1 and ranging to some number **n**, usually less than or equal to 10. Only a handful of features extend beyond ten slots, and a few include a **"-1"** bin to catch entries of "unknown" value. Unknowns are treated inconsistently across the features; this is one more reason to lend credibility to the mixed-origins data conjecture. Sometimes 'unknowns' are assigned the lowest numbered slot, i.e., **"-1"**, or **"0"**, or **"1"** depending on the particular feature's lowest label. Occasionally the highest numbered slot, e.g., **"9"**, or **"10"** is used to encode the unknowns.

Very few features deviated from the pre-encoded ordinal pattern, and needed preprocessing:

- *CAMEO_DEU_2015* : Nominal category with string labels, and **"XX"** bin for unknowns.
- *D19_LETZTER_KAUF_BRANCHE* : Nominal category containing other column names.
- *OST_WEST_KZ* : Nominal binary category distinguishing West from East Germans.
- *CAMEO_DEUG_2015* : Ordinal category with **"X"** default bin for unknowns.
- *CAMEO_INTL_2015* : Ordinal with catch-all default bin **"XX"**, and ambiguous labels.

The extent of preprocessing to deal with these features consisted of two steps:

1. Label Disambiguation: Some features had duplicate slots, **"1"**, and **"1.0"**; **"2",** and **"2.0"**; **"3"**, and **"3.0"**, etc., a mixture of labels of string and numerical types for each bin. These entries needed to be collated to a unique numerical label to disambiguate them.
2. Catch-all vs NaNs: Some features presented non-numerical labels such as **"X"** or **"XX"** to serve the function of catching entries of "unknown" value within the feature. These were encoded with a numerical label, the next available number in order to preserve any order extant in the category.

Imputation: Univariate imputation algorithms work for numerical features and rely on a central tendency measure to deal with a small amount of missing entries. To deal with large amounts of missing data in ordinal features an alternative was sought. In "*Missing Data Imputation for Ordinal Data",* Quintero et al (2018) conclude that *"The Random Selection method is the method with the best performance to treat the type of categorical data of that study"*. Random Selection was found to be superior to the *Most Frequent Value* imputation for categorical data, and better than the *Multiple Imputation by Chained Equations* (MICE) methodology, which takes into account all the features of the dataset.

Rather than imputing with central tendency statistics, or with multivariate techniques, a version of the Random Selection method described by Quintero et al (2018; 3.3), was implemented. This method keeps the feature shapes by using the bins' relative frequencies to distribute randomly imputed values. Even in the context of Multivariate Imputation, Kropko et al (2013) note *"If the data are distributed according to a known distribution, imputing missing values is only a simple matter of drawing from the assumed distribution"* (p. 4); Arguably, the feature distributions are 'known' to the extent that the include a slot for 'unknowns'. Thus, if a requisite in imputing missing entries is to leave the feature distributions unperturbed, then the technique is justifiable.

To deal with missing data in features that deviated from the standard pattern of 10 or less slots, a visualization tool was implemented to examine the effect of random selection on a before and after basis, without actually performing the imputation. The 'before' picture shows the missing data as an arbitrary category, a fake slot with value **"-10"**, for example, shown as a red bar side-by-side with the rest of the data. Next to it, the 'after' picture depicts the "would-be" effect of imputing with the technique. We discriminate between missing entries (to impute), and unknowns such as the **"0"** label below, for feature *ALTERSKATEGORIE_FEIN* — Fig. 2.

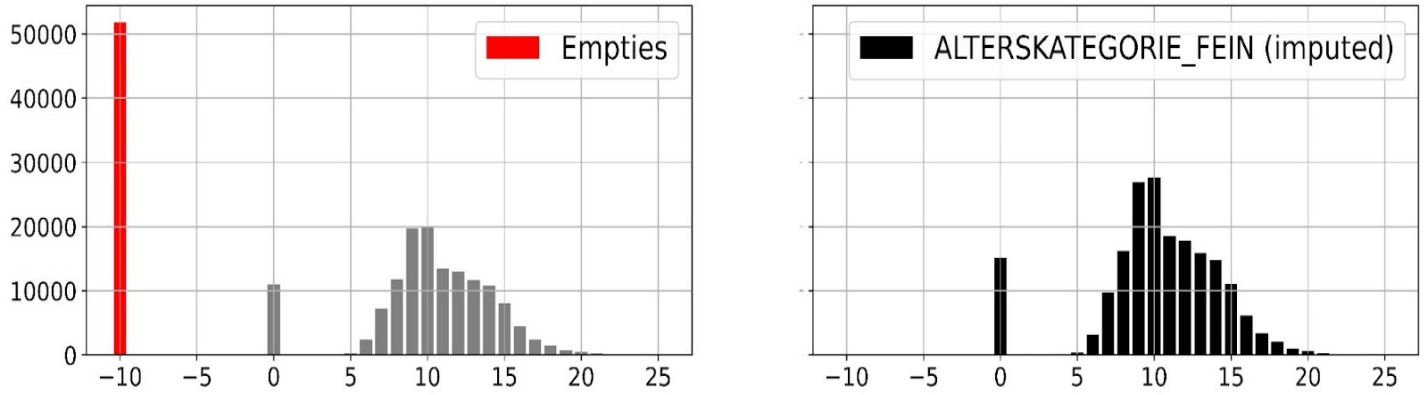

Figure 2 — Random Selection Univariate Imputation.

A handful of atypical features having a much wider range of bins than the rest were imputed using two methods, the mean value imputation, as well as the random selection methodology. Figure 3 shows the results of both strategies on one of these features, *KBA13_ANZAHL_PKW — the number of cars in the zip code*. As expected, given the large amount of missing data, random selection produces a smoother imputation that keeps the feature's distribution intact.

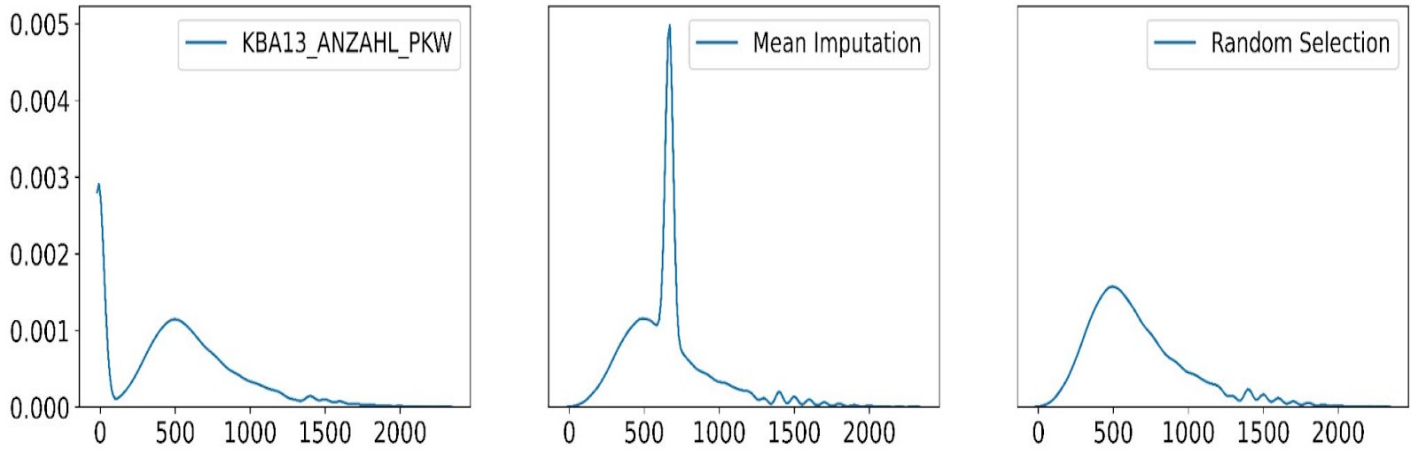

Figure 3 — Mean Imputation vs. Random Selection.

Outliers: The presence of outliers, and the way they are handled can significantly affect the performance of estimators, as they rely on data statistics that get thrown off by bad entries. There are three ways to deal with outliers, Kwak et al (2017 - Introduction), namely, trimming, replacing, and robust techniques. Which method is used depends on whether the distribution is known, and also on the amount of outlier noise. Assuming 'known' distributions, as previously argued, we are left with the consideration of noise alone. Two features are plagued by outliers: *GEBURTSJAHR — year of birth,* with ~50%, and *KBA13_ANZAHL_PKW — the number of cars in the PLZ8* with ~6%. PLZ8 is the German equivalent of the US ZIP+4 regional nomenclature. PLZ8 uses eight digit codes to denote areas of approximately 500 households.

With *GEBURTSJAHR* there is only one outlier point, but the proportion it represents is excessive; almost half of all individuals claim to be born in the year zero. If a sector of the population is unwilling to reveal their true age, and if that sector comprises ~50% of individuals, that may explain the source of this outlier value. *KBA13_ANZAHL_PKW*, on the other hand, contains 11 outlier entries that add up to about 6% of the feature data. But the cause of these outliers is transparent; it is the result of an abrupt change in the granularity of reporting. The amount of cars in any PLZ8 region is precisely reported in unitary increments up to 1,250 cars; beyond this amount, the reports jump to 1,300 units and increase by hundreds all the way up to 2,300.

To deal with these outliers, two different approaches were used. For *GEBURTSJAHR*, a simplifying assumption was made, that the unreported age of individuals follows the same distribution as that of the reported ones. This has the potential to taint any conclusions of the analysis if the true cause of the noise lies with demographics, such as gender or generational differences, i.e., the extent to which women, or men, or older individuals are more, or less resistant to disclosing their age during data collection efforts. The outlier was handled with the same random selection imputation technique that was used to deal with missing entries. Figure 4 shows the feature before the imputation (y-axis clipped to eliminate outlier distortion) above, and after imputation with random selection, below. In absolute terms, the data entries have all been doubled. In relative terms, nothing has happened to the data, except for the removal of noise, while hiding its source and possibly tainting conclusions along demographic arguments.

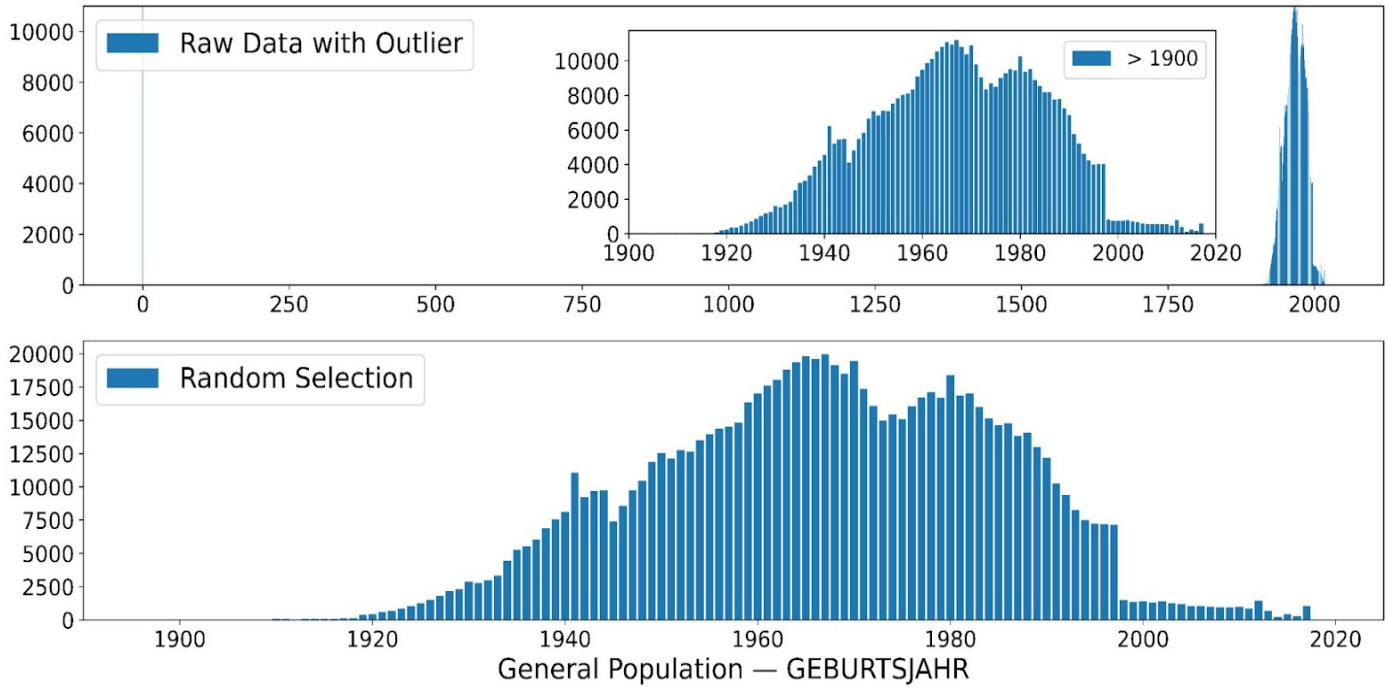

Figure 4 — Handling GEBURTSJAHR Outliers (Random Selection).

With *KBA13_ANZAHL_PKW,* the cause of the outliers (difference in reporting) suggested a fix consisting of reassigning the 11 offending entries in more granular fashion within the right tail where the outliers emerged. Figure 5 shows the right tail outliers above, with refinement, below. To make matters simple, a straight line approximation was used to populate unitary bins from 1,251, to 2,300 using the ~6% of the data contained in the 11 outliers.

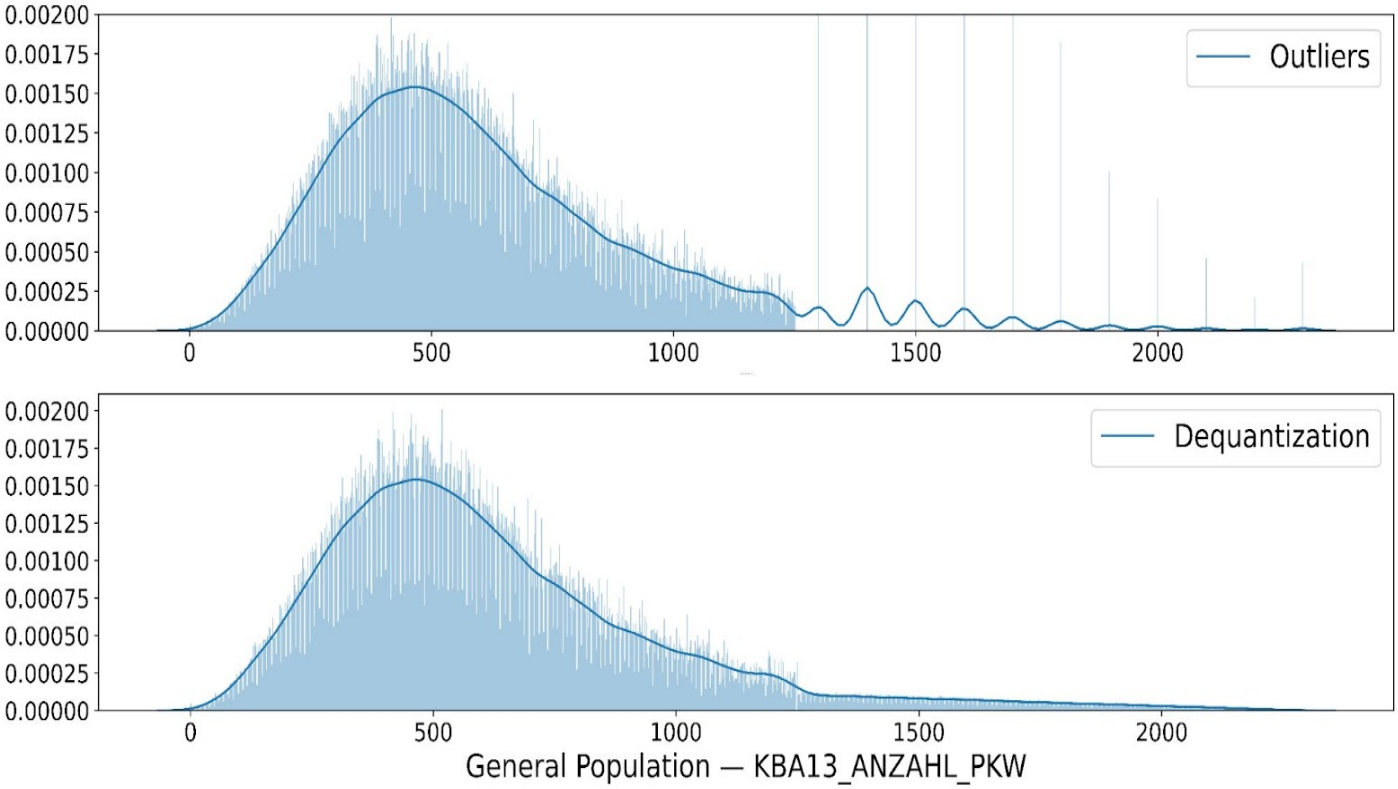

Figure 5 — Handling KBA13_ANZAHL_PKW Outliers (Dequantization).

Visual Analysis: Without ML techniques, it is possible to find features with distributions that differ markedly between the datasets. Using the spreadsheets to decrypt the meaning of the feature names, a picture of how customers stand out from the population begins to emerge. Unfortunately, not all of the feature names are described, but some can be translated, such as *GEMEINDETYP* — type of community. Others, such as *HH_DELTA_FLAG* are anyone's guess. Features whose names could not be decoded had to be ignored in this part of the analysis.

The datasets were compared side by side on a feature by feature, and slot by slot basis using visualizations of the kind shown below — Fig. 6. Feature slots where customers exhibited overrepresentation, or underrepresentation by more than some tolerated level, say 10%, were flagged as interesting for further examination.

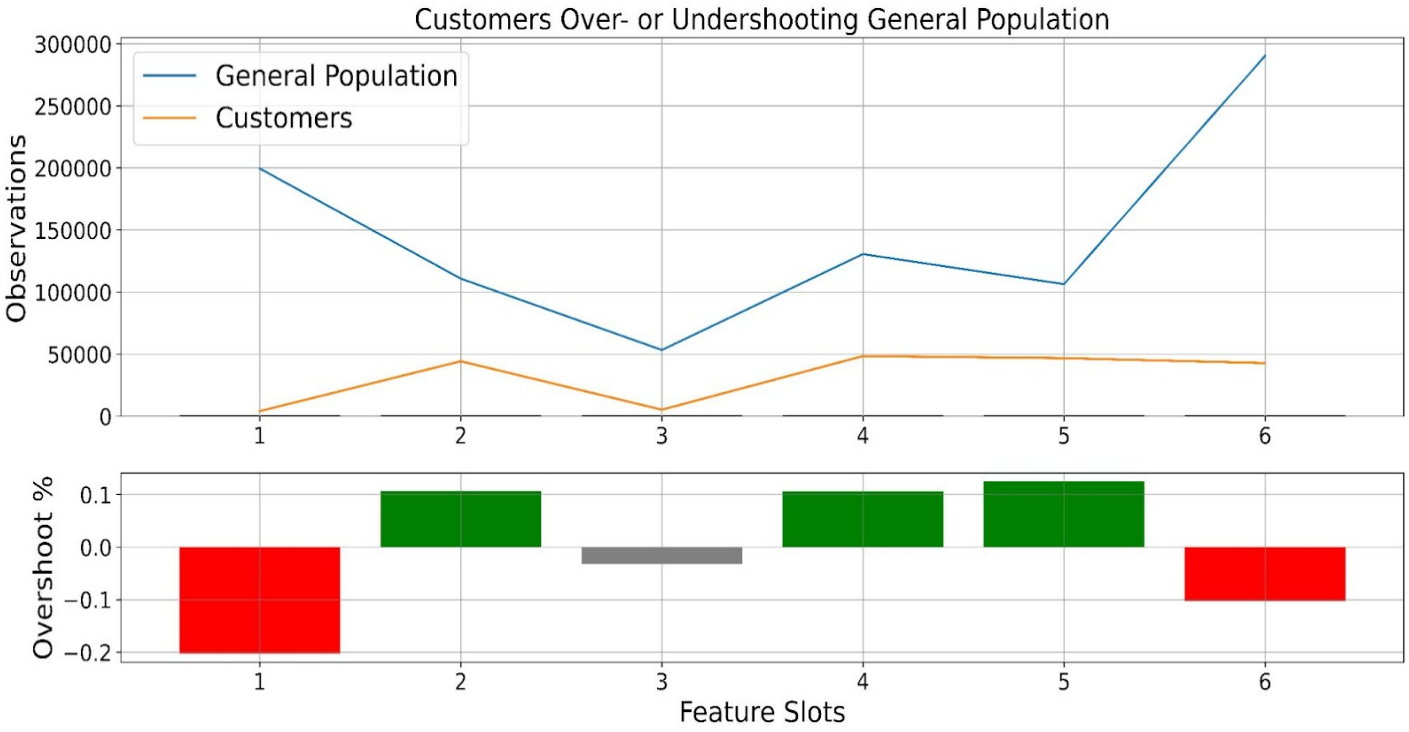

Figure 6 — General Population vs. Customers.

Enough acquaintance with the data through interactive visualizations reveals a profile for customers as individuals that tend to be shoppers of diverse products, but not online; they are financially unremarkable; they tend to save and invest, but not to prepare. They are fans of nature, with high income, residing longer than 10 years in areas of mostly 1-2 family homes, without children living at home, and not very interested in moving. They are environmentalists, but have multiple cars in the household. They are mostly old-fashioned (traditional minded) older males from the west of Germany, around 60 years of age, not single, top earners, with Germanic sounding names, dominant traits, critically minded, and not very social.

N.B. If an unwillingness to respond truthfully to questions regarding age exists in a large segment of the population (~50%), then this customer's profile conceals that information. For example, if roughly half of participants were women in their 40's, or men in their 50s, or environmentalist teenagers, or anyone who would balk at the prospect of revealing something to intimate, the above customer profile will hide all that and make every customer look like another segment of the population, the segment willing to answer truthfully. This is because of our handling of the 50% imputation of the "year of birth" feature — *GEBURTSJAHR*. A quick peek at the histogram shows two peaks, one at 1965, and another at 1980, biasing the customers' profile towards the tendencies of individuals in their late 50's and early 40's respectively.

Unsupervised Learning — Customer Segmentation

Keeping in mind our ultimate goal of improving customer acquisition using targeted marketing, we now turn to unsupervised learning ML techniques; Chinedu et al (2015) remark *"The importance of customer segmentation include, inter alia, the ability of a business to customize market programs that will be suitable for each of its customer segments" (p.40).* Customer Segmentation aims at recasting the customers observations into a few segments. *"Each segment comprises customers who share similar market characteristics" (Abstract).* Chinedu et al (2015).

Principal Component Analysis: It is easier to form clusters if we start with fewer dimensions. In high dimensionality cases, such as ours, with 366 features, a reduction step takes place before clustering. Principal Component Analysis (PCA) is a dimensionality reduction technique that projects the original data features onto a space of uncorrelated new features (components) that successively maximize retained variance at the cost of dropping a small number of features. Holland (2019) notes *"several criteria have been proposed for determining how many PCs should be investigated and how many should be ignored. One common criteria is to ignore principal components at the point at which the next PC offers little increase in the total variance explained. A second criteria is to include all those PCs up to a predetermined total percent variance explained, **such as 90%**" (p. 3)*.

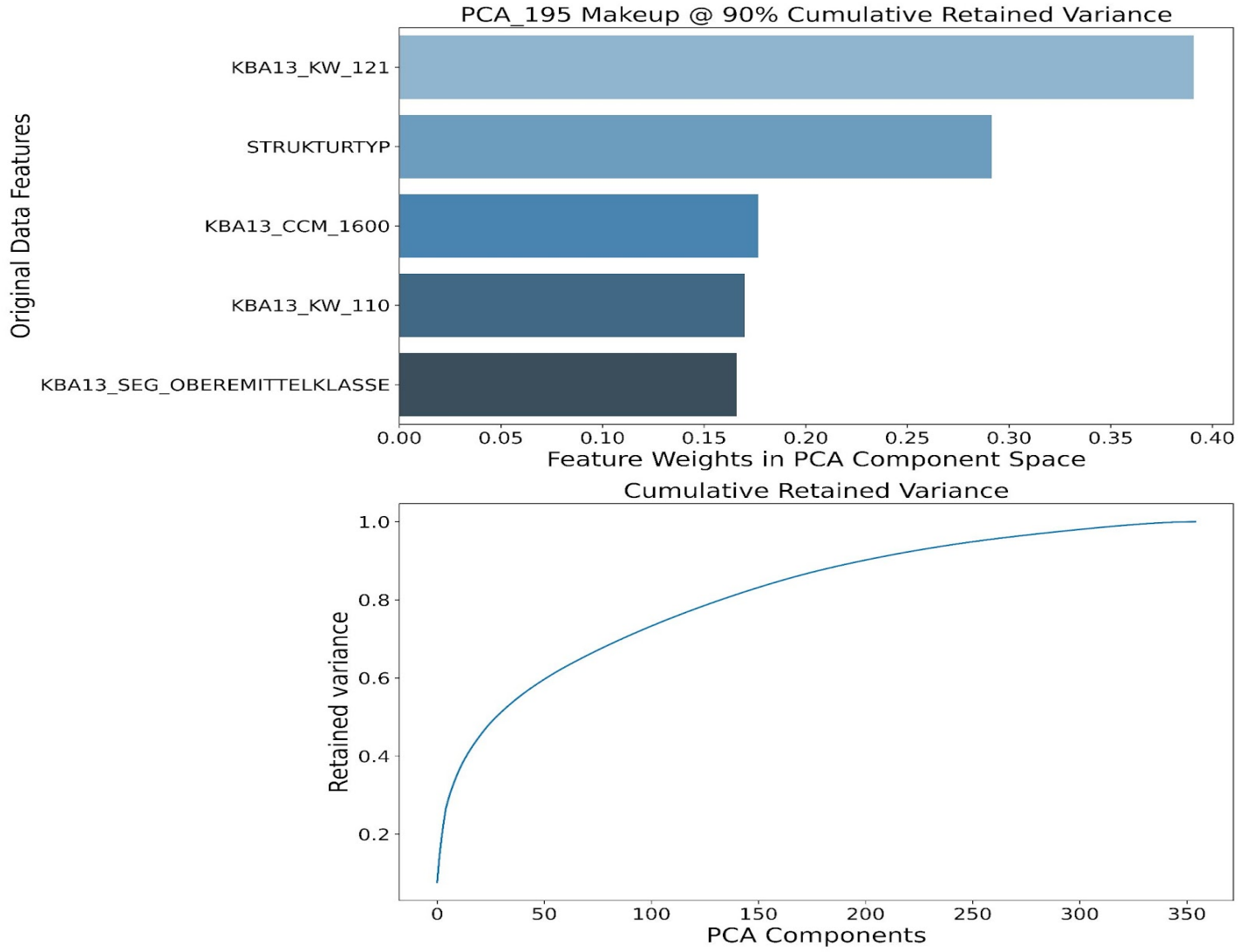

Figure 7 — PCA Component Features and Retained Variance.

Using PCA the features were projected onto a 90% retained variance space, using 195 principal components instead of the original 366 columns, thus trading off 10% of the information contained in the original variables to attain a 50% reduction in dimensionality. The linear combination of features making up each principal component as well as the cumulative amount of variance retained by each, were made visible interactively. Figure 7 shows the PCA component where the cutoff was made, *PCA_195*, along with the top five features in its make up.

K-Means Clustering: Next, the complexity of the problem was further reduced by segmentation (clustering). The idea behind clustering is to turn the 891,221 observations from the general population into a smaller number, say K, of so-called clusters of points. If each observation can be thought of as a point, then each cluster, as the name indicates, is a conglomerate of points, a super-point that for all intents and purposes can be viewed as a good representative of the observations near it. *"K-Means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem … through a certain number of clusters fixed a priori" (p. 91)*, Trupti et al (2013). Without a priori knowledge of the number of clusters to use, several segmentation attempts were made by varying the number of clusters **k**, to examine how thinly or densely populated the resulting clusters were. Figure 8 shows the results of plotting the average cluster centroid distances against the values of **k** attempted.

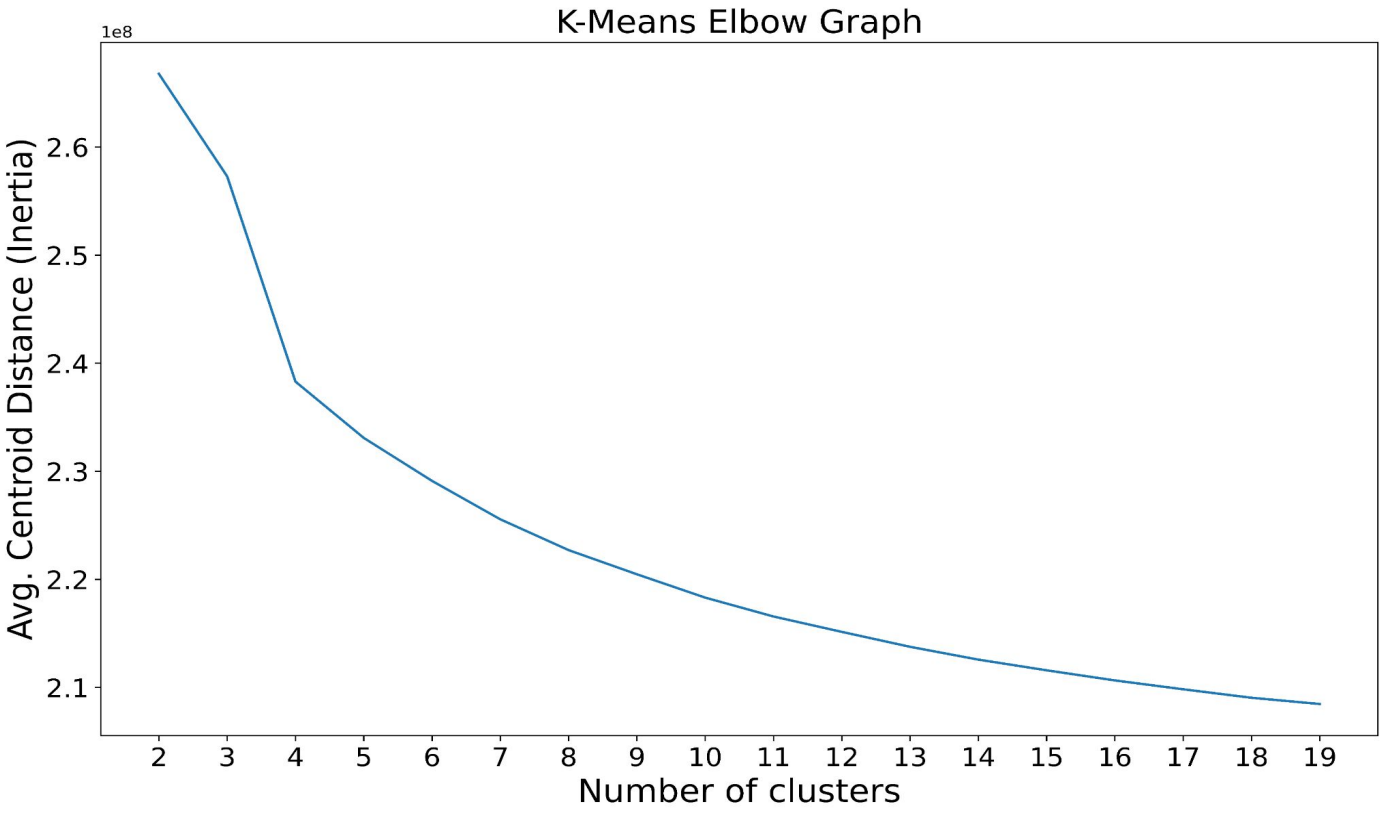

Figure 8 — K-Means Average Centroid Distance vs Number of Clusters.

With K-Means, the value of **"k"** is chosen so that the average centroid distance is low, i.e., the clusters are packed. *"The oldest method for determining the true number of clusters in a data set is inelegantly called the elbow method. … at some value for K the cost drops dramatically, and after that it reaches a plateau when you increase it further. This is the K value you want" (p.92)*, Trupti et al (2013). Following this advice, a choice was made at **"k=4"** for our analysis.

Cluster Analysis: The PCA and K-Means models were fitted using the general population, and then used to transform the customers dataset. Figure 9 compares how the observations were assigned to one of four cluster labels in absolute terms (number of observations), in relative terms (densities), and most importantly, the bottom subplot shows the degree to which each cluster over represents (green) or under represents (red) the customers. To guide a targeted marketing campaign, we need to investigate what these clusters mean in terms of demographic traits, i.e., in terms of original features now obscured behind the PCA components and clusters. We are particularly interested in clusters 1 and 0 which highly overrepresent the customers.

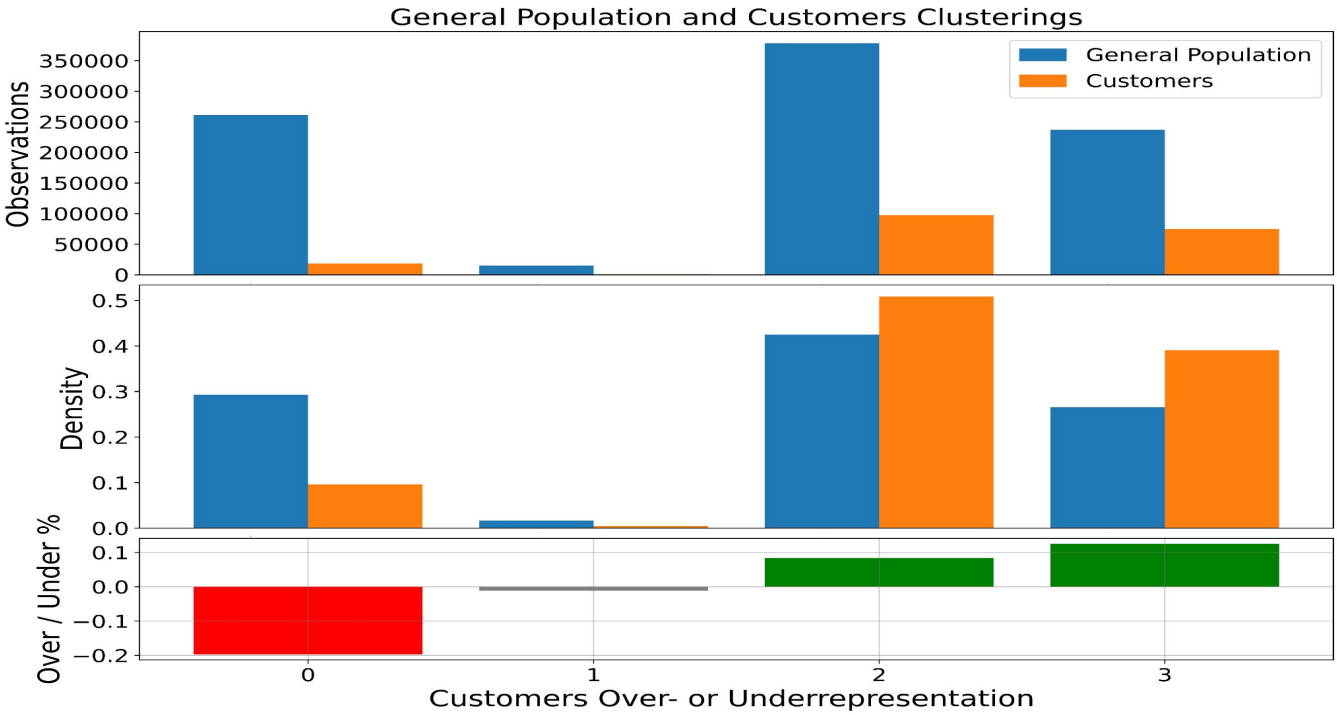

Figure 9 — Cluster Observations, Densities, and Over- Under-Representation.

Getting back the original features is a two-step process. First, the coordinates of the clusters in 195-dimensional PCA space are examined. These coordinates get progressively closer to the origin for higher components, so the search for meaning can be narrowed to the first 20 or so PCA axes. Figure 10 shows a heatmap using only the first 20 PCA components. The stronger shading reveals the axes with the largest cluster projections; so for example, for cluster `c_0` the largest coordinates project onto `PCA_4`, `PCA_3`, `PCA_1`, `PCA_14`, and `PCA_6` in that order. Further projections can be extracted, but their impact drops quickly in relevance. Four our study, only the top five projections were considered, for the sake of expediency.

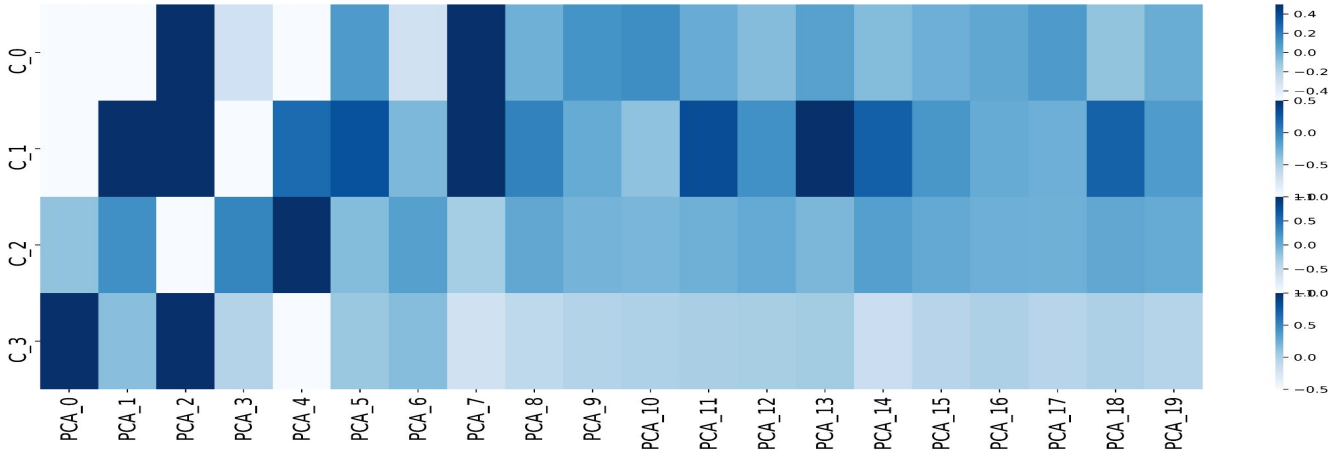

Figure 10 — Cluster PCA Components Heatmap.

The second step in getting back the original features consists of breaking down the main PCA cluster projections into their constituent features by weight, as was done in the previous PCA analysis depicted in Figure 7. These two steps can be combined and automated to dissect the clusters as shown in the abridged output snippet below. The clusters are sorted by over (green) or under (red) representation with customers; their main PCA axes are listed, and along with them, the main features. It is then a simple matter of looking up the meaning of these features in the spreadsheets, and examining the feature slots over or under representation by customers, as was shown in Figure 6, to arrive at the desired meaning of the segmentation exercise.

```
C_3 Overrepresentation by 12.51%
    PCA_0 : ['LP_STATUS_FEIN +12.35', 'MOBI_REGIO +12.28', ...]
    PCA_2 : ['ONLINE_AFFINITAET +14.4', 'D19_GESAMT_ANZ_24 +14.28', ...]
    PCA_6 : ['SEMIO_KAEM +34.44', 'ANREDE_KZ +33.28', 'SEMIO_VERT -32.39',...]
    PCA_1 : ['KBA05_SEG6 +17.89', 'KBA05_KRSOBER +16.33', ...]
    PCA_5 : ['KBA13_KW_0_60 +21.01', 'KBA13_KW_61_120 -20.9', ...]
C_2 Overrepresentation by 8.38%
    PCA_4 : ['FINANZ_ANLEGER +18.51', 'FINANZ_SPARER +18.27', ...]
    PCA_3 : ['KBA13_HERST_BMW_BENZ +19.74', 'KBA13_MERCEDES +17.04', ...]
    PCA_1 : ['KBA05_SEG6 +17.89', 'KBA05_KRSOBER +16.33', ...]
    PCA_14: ['D19_BANKEN_ANZ_12 +25.09', 'D19_BANKEN_ANZ_24 +24.35', ...]
    PCA_6 : ['SEMIO_KAEM +34.44', 'ANREDE_KZ +33.28', 'SEMIO_VERT -32.39', ...]
C_0 Underrepresentation by -19.71%
    PCA_2 : ['ONLINE_AFFINITAET +14.4', 'D19_GESAMT_ANZ_24 +14.28', ...]
    PCA_7 : ['VERS_TYP +29.37', 'HEALTH_TYP +29.23',  ...]
    PCA_10: ['LP_LEBENSPHASE_GROB +16.89', 'LP_LEBENSPHASE_FEIN +16.73', ...]
    PCA_9 : ['KBA13_KMH_140_210 +18.75', 'KBA13_CCM_1401_2500 +18.25',  ...]
    PCA_5 : ['KBA13_KW_0_60 +21.01', 'KBA13_KW_61_120 -20.9', ...]
C_1 Underrepresentation by -1.17%
    PCA_1 : ['KBA05_SEG6 +17.89', 'KBA05_KRSOBER +16.33',  ...]
    PCA_2 : ['ONLINE_AFFINITAET +14.4', 'D19_GESAMT_ANZ_24 +14.28',  ...]
    PCA_7 : ['VERS_TYP +29.37', 'HEALTH_TYP +29.23', 'KOMBIALTER -28.8', ...]
    PCA_13: ['KBA13_HERST_AUDI_VW +20.48', 'KBA13_VW +19.22', ...]
    PCA_11: ['LP_FAMILIE_FEIN +17.1', 'LP_FAMILIE_GROB +16.58', ...]
```

This exercise revealed that the population can be grossly considered to be comprised of four types of individuals, two of which (Clusters 1 and 0) are worthwhile pursuing with targeted marketing campaigns. A brief description of all the clusters follows:

- *Cluster 3 — Wealthy Older West German Males:* This group has the highest customer overrepresentation and consists of men who grew up in post-war West Germany. They are sedentary, culturally minded high income earners, with unremarkable online presence. Some are new homeowners and live in areas where there are some upper class cars, such as BMWs, and some trailers, but not many cars built between 2000 and 2003.
- *Cluster 2 — Money Savvy Wealthy Males:* Also strongly overrepresented with customers. These are men who save or invest their money, rather than spending it in the home, or being prepared; they are not known for online-banking. They are culturally minded, and live in areas with a very high share of newly built cars as well as upper class cars, such as BMWs and Mercedes Benz, but not too many 5-seaters.
- *Cluster 0 — Power Couples & Retirees:* Hugely underrepresented with customers. These are either high-earning couples or retiring single high-earners that live in multigenerational households in areas with a high share of newly-built cars. They have Germanic last names and are very demanding shoppers, but have an unremarkable online presence.
- Cluster 1 — Older Couples & Households: Also underrepresented with customers, this group consists mostly of people who are not single and who are living in multigenerational households. Nothing is known of their wealth status, but their online presence is unremarkable and they are very demanding shoppers. They have Germanic last names and grew up in post-war Germany.

We should note that these observations are based on incomplete information. Many features were ignored when we decided to analyze only the top five features per PCA component. Furthermore, of the features considered for analysis, several could not be decoded from the spreadsheets, and had to be ignored as well. But if we had to venture a guess as to where the firm ought to focus its marketing efforts, it would be on culturally minded men living in areas with a high share of new and upper class vehicles. Men who are either money savvy from anywhere in Germany, or wealthy and older "grandpa'' types from the west of the country. Conversely, not much can be gained by advertising to individuals who live in multigenerational households, regardless of wealth or share of new and upper class cars in their area.

Supervised Learning — Classifier

The next part of the project — supervised learning, entailed training a classifier model with the mailout data, and fine-tuning some of its parameters in order to make predictions that would eventually be submitted into a Kaggle Competition. All prior data wrangling steps were carried out as was done for the general population and customers in the first part of the project.

Metric Selection: A quick inspection of the mailout labels shows an imbalance in the data; the presence of customers is rare — Fig. 11. This eliminates accuracy as a metric for our study. Indeed, with such high imbalance, any model predicting "non-customer" for every observation would be highly accurate, but of little use. *"Considering the case of imbalanced data assessment; Accuracy places more weight on the common classes than on rare classes, which makes it difficult for a classifier to perform well on the rare classes"* Bekkar et al (2013 p.28).

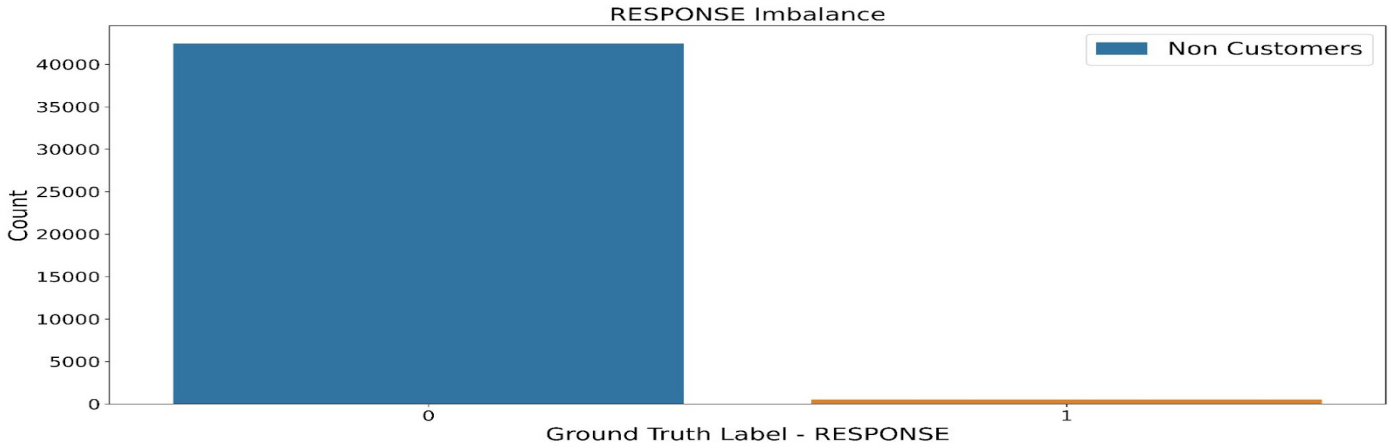

Figure 11 — Mailout Data, Labels Imbalance.

Rather than accuracy, a more fundamental evaluation metric is needed; one that accounts for false positives and false negatives in a sensible way for the targeted marketing problem at hand. We are more inclined to tolerate false positives, i.e., marketing to non-customers at the risk of annoying them slightly. False negatives, on the other hand, are a serious problem since they translate directly into missing out on the potential revenue from misidentified true customers. This means that we care more about *Recall* or *Sensitivity*, than we do about *Precision*.

Recall, or *True Positive Rate* (TPR), captures the fraction of correctly predicted customers to the total pool of customers,

Recall = true positives/(true positives + false negatives).

Similarly, *False Positive Rate* (FPR), captures the fraction of incorrect predictions to the total pool of non-customers,

False Positive Rate (FPR) = *false positives/(false positives + true negatives)*.

The Receiver Operating Characteristic — ROC, is a curve that gives TPR as a function of FPR. *"The more inclined the curve is toward the upper left corner, the better is the classifier's ability to discriminate between positive and negative class"* Bekkar et al (2013 p.30).

By measuring the ROC Area Under Curve (AUC), we have a metric that summarizes the performance of classifiers as a single number. The AUC provides us with a scoring mechanism to tell how well a model distinguishes customers from non-customers; therefore, ROC AUC was picked as our metric of choice for model selection where we compared several models, as well as for parameter tuning, where we compare a model against a potentially better version of itself.

Model Selection: Several classifiers were tried, their ROC curves plotted, and the corresponding ROC AUC scores computed in order to discern the most promising model for subsequent tuning. The list of the models attempted, their corresponding ROC AUC scores, and their ROC curves are shown together in Figure 12. We explored *Logistic Regression*, as well as ensemble techniques like *Random Forest*, *Adaptive Boosting*, *Gradient Boosting Machine* and the *XGBoost* variant.

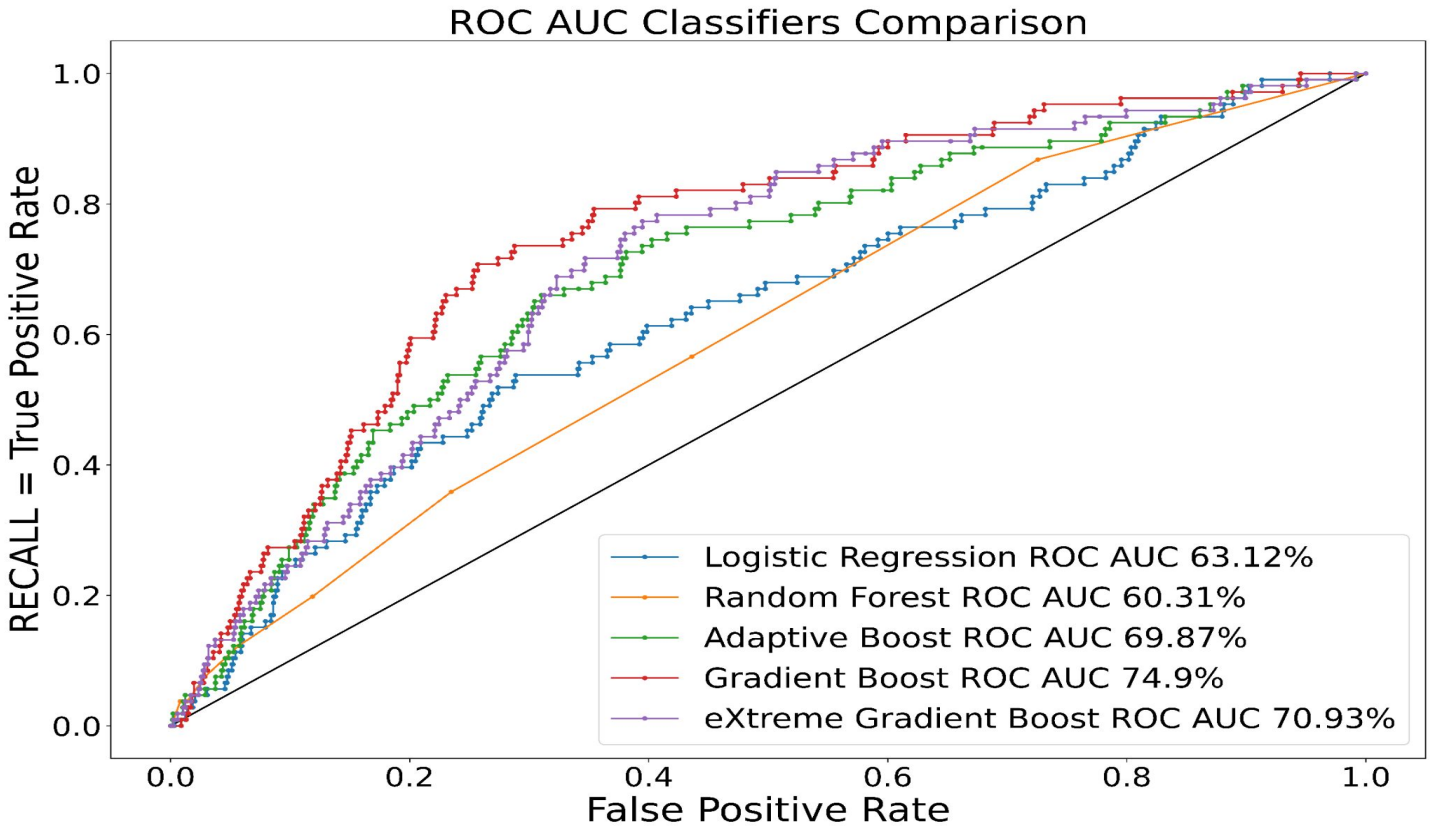

Figure 12 — ROC Curve and ROC AUC Metric for Classifier Model Selection.

The table in Figure 13 serves as guidance for the meaning of the AUC values, Allaire (2006).

AUC Value	Model performance
0.5 - 0.6	Poor
0.6 – 0.7	Fair
0.7 – 0.8	Good
0.8 – 0.9	Very Good
0.9 – 1.0	Excellent

Figure 13 — AUC Value Interpretation.

Guided by the interpretations listed on the table in Figure 13, all the models attained fair to good performance, with the Gradient Boosting algorithm emerging as the clear choice for parameter tuning and predicting. Bekkar et al (2013 p.30) note that *"Several ROC curves can be represented in the same space to compare the results of different models ... a curve dominating the others and modeling associated with the dominant curve is considered more efficient"* . The Gradient Boost ROC curve can be seen to be markedly above the rest, showing a trajectory nearest the upper left corner, and with an AUC score of 74.9%, it is several percentage points ahead of the next best classifier model (Fig. 12).

Parameter Tuning: The parameters of the Gradient Boosting model can be divided into 3 categories: Tree-Specific Parameters, Boosting Parameters, and other miscellaneous parameters for overall functioning. We focused first on a few of the most important tree-specific, and boosting parameters, namely, learning rate (*learning_rate*), number of trees (*n_estimators*), tree depth (*max_depth*), and the number of randomly selected features to consider while searching for a split (*max_features*).

Boosting is a sequential process that may, in some cases, tend to overfit the training data. To avoid overfitting, several models were trained by varying the parameters of interest over a range of values in isolation. Then the ROC AUC score was computed for each trial value of the parameter, and the resulting train and validation (test) learning curves were compiled. The curves were observed to search for ranges where the model's learning curves' gap between the train and test results weren't too wide (Fig. 14). The *learning_rate* parameter, for example, on the top-left graph, shows a very wide gap using a value of 0.1; there the ROC AUC difference between the learning curves, ~17%, indicates that the model is overfitting the training data, and thus not generalizing well.

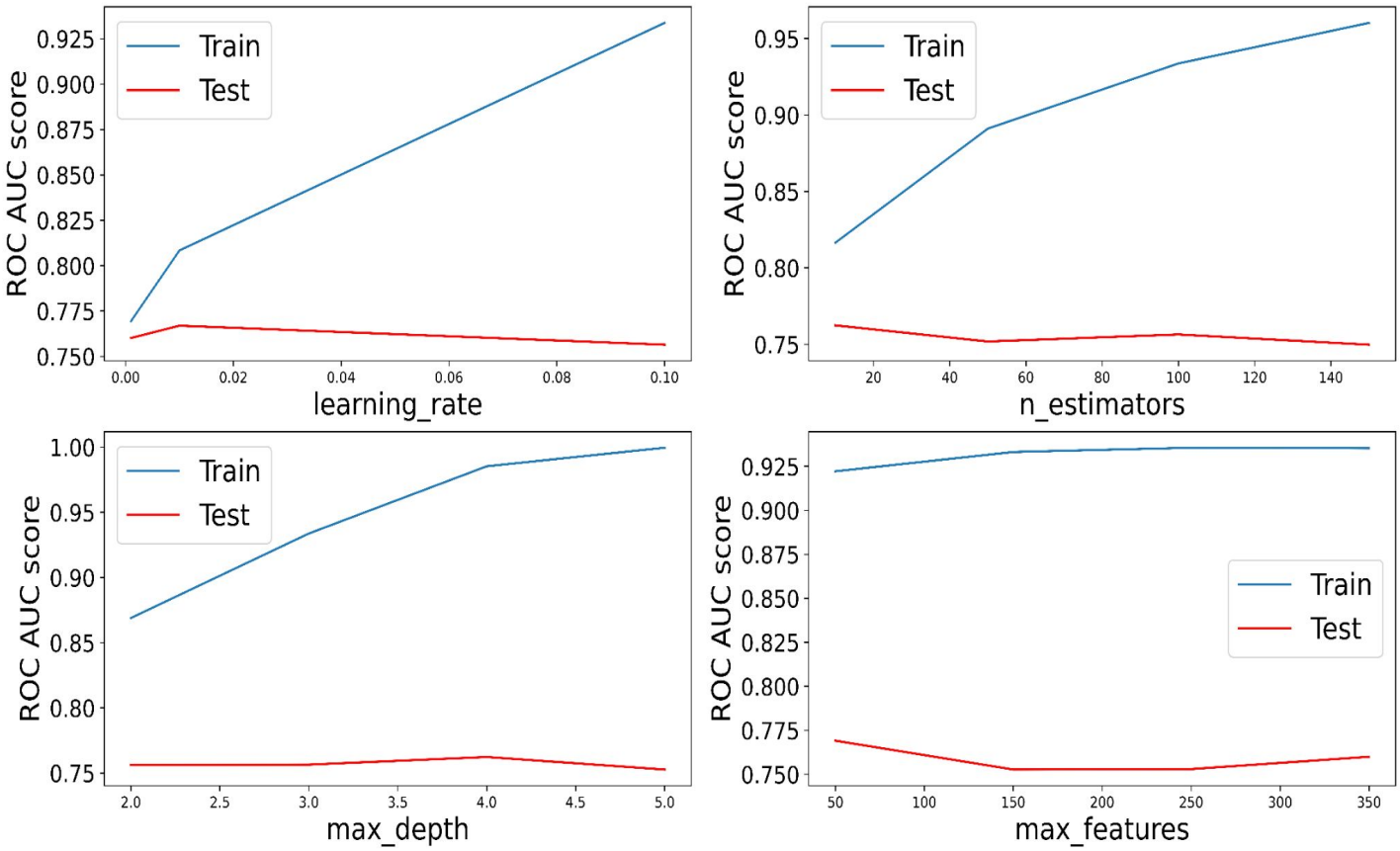

Figure 14 — Learning Curves, Train and Validation(Test), for Several Isolated Parameters.

The actual tuning was carried out with SK-learn's *GridSearchCV* library function, fitting 5 folds to cross validate each of the parameter value combinations. The parameter search resulted in marginal improvement, with AUC reaching 76.44% using the following parameter values: *{'learning_rate': 0.01, 'max_depth': 4, 'max_features': 225, 'n_estimators': 50}*.

Two further parameters were tuned: the fraction of observations to be selected for each tree (*subsample*), and the minimum number of observations which are required in a node to be considered for splitting (*min_samples_split*). Tuning again by fitting 5 folds to cross validate resulted in an improved AUC reaching xx.xx%. With our Gradient Booster tuned, we measured improvement in predicting power over our baseline model, 78.55% vs 76.5%. Correspondingly, our best submission to the kaggle competition was scored at merely 79.61%. Details of the tuning, and output from cross validations are available in the notebook *Capstone Arvato.ipynb*.

Conclusion: Unsupervised and supervised ML techniques were implemented to identify customers from demographic data, for a targeted marketing campaign. There were challenges with the data: A significant number of rows had more than 70% blank entries, there were jumps in the amounts of missing data, and the meaning of many features was missing from the ancillary spreadsheets. Additionally, some features needed extra wrangling steps to disambiguate their slot labels, to impute their missing entries, and to handle their outliers. One feature in particular, *GEBURTSJAHR* or "year of birth" contained 50% of its entries as noise — roughly half of all individuals were registered as being born in the year zero!

After exploring and preprocessing the data, the distributions were compared to focus on features and slots that over or underrepresented the customers. The profile that emerged describes customers, among other things, as wealthy and sedentary nature lovers, mostly West German old males living without children at home. An important remark was made with respect to the potential concealment of meaningful customer attributes in this study. The combination multiple unexplained features, as well as the handling of the "year of birth" noise will likely result in incomplete and tainted conclusions.

Using unsupervised learning ML techniques we first reduced dimensionality by 50% with Principal Component Analysis, at the cost of 10% of the information. Subsequent K-Means Segmentation using four clusters revealed a more refined profile for the customers. Briefly, they fall into two broad segments, namely, wealthy older West German males, and money-savvy wealthy males.

Finally, using supervised learning ML techniques, a Gradient Boosting Classifier was selected and tuned with the ROC AUC metric, and attained a score of 79.6% in the Kaggle Competition.

There is definitely room for improvement in this analysis, besides the obvious need for a complete description of all columns in the datasets. As far as the preprocessing piece, there were a few features showing inner structure; they could be reengineered as multiple separate variables with independent information. But more importantly, given the excessive amount of missing data in this project, it is probably here that that sweet spot can be found for improving both the unsupervised as well as the supervised results. Multivariate imputation techniques, such as MICE should be explored to see if indeed the random selection approach we opted to use is the better one.

Overall, the project was fun; a lot was learned not only with regards to ML techniques, but also about the tools of the trade as applied to a real-life problem. I would recommend the course, and the project to anyone genuinely curious about Machine Learning.

References:

Maryuri Quintero, Aera LeBoulluec, (2018). Missing Data Imputation for Ordinal Data. https://www.researchgate.net/publication/326435546_Missing_Data_Imputation_for_Ordinal_Data

Lokesh, (2021). Dealing with Missing Values for Data Science Beginners. https://www.analyticsvidhya.com/blog/2021/10/guide-to-deal-with-missing-values/

Jonathan Kropko, Ben Goodrich, Andrew Gelman & Jennifer Hill, (2013). Multiple Imputation for Continuous and Categorical Data:Comparing Joint and Conditional Approaches, page 2., http://www.stat.columbia.edu/~gelman/research/published/MI_manuscript_RR.pdf

Sang Kyu Kwak and Jong Hae Kim, (2017). Statistical data preparation: management of missing values and outliers, Korean journal of Anesthesiology. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5548942/

Chinedu Pascal Ezenkwu, Simeon Ozuomba, Constance Kalu, (2015). Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services; International Journal of Advanced Research in Artificial Intelligence. https://www.researchgate.net/publication/282862569_Application_of_K-Means_Algorithm_for_Efficient_Customer_Segmentation_A_Strategy_for_Targeted_Customer_Services

Steven M. Holland, (2019). Principal Components Analysis. http://strata.uga.edu/software/pdf/pcaTutorial.pdf

Trupti M. Kodinariya, Prashant R. Makwana, (2013). Review on determining number of Cluster in K-Means Clustering, International Journal of Advance Research in Computer Science and Management Studies, https://www.researchgate.net/profile/Trupti-Kodinariya/publication/313554124_Review_on_Determining_of_Cluster_in_K-means_Clustering/links/5789fda408ae59aa667931d2/Review-on-Determining-of-Cluster-in-K-means-Clustering.pdf

Mohamed Bekkar, Dr.Hassiba Kheliouane Djemaa, Dr.Taklit Akrouf Alitouche, (2013). Evaluation Measures for Models Assessment over Imbalanced Data Sets, Journal of Information Engineering and Applications. https://eva.fing.edu.uy/pluginfile.php/69453/mod_resource/content/1/7633-10048-1-PB.pdf

Allaire JF, (2006). Introduction à l'analyse ROC Receiver Operating Characteristic, Centre de recherche Institut Phillipe-Pinel de Montréal, école d'été.