UDACITY

MACHINE LEARNING ENGINEER NANODEGREE

Capstone Proposal

Customer Segmentation – Arvato Financial Solutions

Ozayr Moorad

April 22nd 2020

1. DEFINITION

1.1 PROJECT OVERVIEW

1.1.1 DOMAIN BACKGROUND

Arvato Bertelsmann is a global services company founded in 1996 situated in Gutersloh Germany[Ostrowski, J. (2016). "Arvato baut aus". Neue Westfälische (in German).]. It provides services in customer support, logistics, finance and IT. The domain of interest for this project is the financial solutions services that Arvato provides, more specifically data driven targeted marketing. The client, a mail order company, requires information as to what type of demographic group their customers fall into. They wish to use this information to concentrate their marketing campaigns to ensure maximum customer acquisition. Hence the goal of this project will be to utilize the provided data augmented with machine learning techniques to uncover patterns that a human analyst might miss.

1.1.2 THE CLIENT

Not much is known about the client except that it is a mail order company that's wants to ensure datapoints are being used to facilitate the decision-making process. I thought it apt to add a few points from the definition of a mail order company that I had come across during my research [https://www.referenceforbusiness.com/small/Inc-Mail/Mail-Order-Business.html]

- Mail-order businesses date back to pre-Revolutionary War days, when gardeners and farmers ordered seeds through catalogues.
- Historically, mail-order businesses became successful because they offered a wider variety of goods than could be found in local retail outlets.
- goods purchased through the mail were often less expensive than those available locally.
- mail-order houses, blessed with the capacity to maintain far larger inventories than many of their retail competitors, could afford to offer more sizable discounts.
- Indeed, for consumers in remote rural sections of the country, their isolation from commercial centres made catalogue or mail-order shopping a necessity.
- Finally, individuals pursuing a hobby or special interest were more likely to locate those hard-to-find items in a specialty catalogue than in a store.

1.1.3 MOTIVATION

One of my motivations for choosing this project was to test the hours of data science reading that I have done over the past few months ever since the beginning of my journey into Machine learning, I believe in order to be a good machine learning engineer one has to have good data science skills. This project would also give me the ability to put my insight extraction and philosophical skills to the test on something that has real world implications as well gain additional skills, knowledge and experience.

1.2.1 DATASETS AND INPUTS

The following information on the datasets to be used are taken as if from the project workspace.

There are four data files associated with this project:

- Udacity_AZDIAS_052018.csv: Demographics data for the general population of Germany; 891 211 persons (rows) x 366 features (columns).
- Udacity_CUSTOMERS_052018.csv: Demographics data for customers of a mail-order company; 191 652 persons (rows) x 369 features (columns).
- Udacity_MAILOUT_052018_TRAIN.csv: Demographics data for individuals who were targets of a marketing campaign; 42 982 persons (rows) x 367 (columns).
- Udacity_MAILOUT_052018_TEST.csv: Demographics data for individuals who were targets of a marketing campaign; 42 833 persons (rows) x 366 (columns).

Provided along with the data set are 2 additional files. Viz:

- DIAS Information Levels - Attributes 2017.xlsx

And is described as a top-level list of attributes and descriptions, organized by informational category.

- DIAS Attributes - Values 2017.xlsx

is a detailed mapping of data values for each feature in alphabetical order.

These files will be used to gain an understanding of the features which will help with the cleaning as well as the interpretation process.

1.2 PROBLEM STATEMENT

The question we seek to answer as put forth by Timo Reis from Arvato is “How can the mail-order company acquire new clients more efficiently?”.

For the first part of the project i.e. customer segmentation, one set of solutions we may investigate or implement is the utilization of machine learning models that will allow us to uncover patterns and insights from the data provided without needing some sort of definitive labels.

- We may use unsupervised learning techniques such as clustering to cluster the population and customers into segments giving us an idea of who or what type of people our customers are.
- Since we have a high number or features available to us, we will need to pre-process, select appropriate features and analyse the dataset such that we achieve the best and most interpretable results for the task of customer segmentation.

The second part is well suited for supervised machine learning, where we are trying to predict who will respond to a mail order campaign given historical data of who had responded to a previous mail order campaign.

- This could be seen as a separate predictive modelling task or can be augmented from insights gained or analysis done on the previous task.
- There are host of algorithms that we could use for this task such as tree-based models or even deep learning models.
- This part will also require analysis to pre-process and select features that will enhance the model's predictability and interpretability.

1.3 METRICS

I will be using 2 main metrics to evaluate models trained, area under the receiver operating characteristic (AUC ROC) curve and area under the precision recall curve (AUC PR).

- AUC ROC: This metric is commonly used to evaluate binary classification. It primarily gives us an indication of how well the model can distinguish between classes. An AUC value close to 1 tells us the model simultaneously classifies each class correctly. A value close to 0 might indicate the model is actually predicting classes opposite to the correct class simultaneously.
- AUC PR: The main difference between the above and this metric is how true negatives are handled. It allows us to better quantify the difference between a model that correctly predicts 10 datapoints out of 100 verses a model that predicts correctly 10 datapoints out of a 1000. This can be very handy when dealing with datasets that have large imbalances in the classes in that we can get a more accurate evaluation of the skill of a model. Basically, since the positive class is outweighed significantly by the negative class this metric will provide useful diagnostic information.

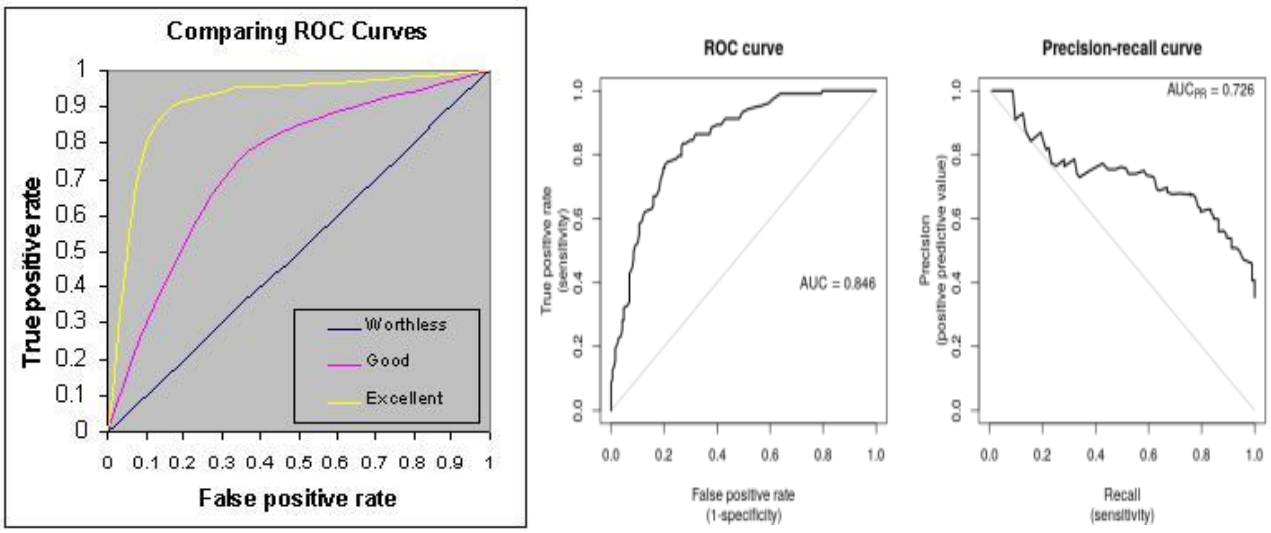

Figure 1: ROC AUC and AUC PR curves

2. ANALYSIS

2.1 DATA EXPLORATION

The main data set for this project is the AZDIAS dataset which consists of a sample of the population of Germany. The dataset consists of 366 features describing demographic, geographical, transactional and transportational information. The data uses approximately 2.4+ GBs of memory. This is quite a significant amount and needed to be optimized which was done by simply recasting data types to represent succinctly the data contained. The optimization brought down the memory usage by more than 50% this enabled faster processing across all steps.

2.1.1 PRE-ANALYSIS

As a pre analysis step I carried out the following:

- Analysed the columns that were not present in the AZDIAS data set but present in the Customer dataset
- Developed an intuition as to what I expect to find based on research of the definition of a mail order company, which I will use as a sort of sudo metric to evaluate the results of the end segmentation.
- Formulated the task from a supervised learning perspective to extract quick insights

There are 3 columns exclusive to the Customers dataset viz.
PRODUCT_GROUP,ONLINE_PURCHASE,CUSTOMER_GROUP

Figure 2: Customer exclusive columns

It is clear to see that many of the customers do not buy online and that cosmetics are bought more online then food, this makes sense.

2.1.2 INTUITION

I had gathered a few questions that I would have liked to be answered to get a better context into the problem I was trying to solve, after answering the questions based on research I had come up with the following conclusion:

So mail order companies are old business structures/types, one thing this tells me is older people would have a sort of comfort toward it, we might get a hint of this from noticing that very few of the customers make online purchases .If you are money savvy you would generally lean toward mail order since it means cheaper. People in rural areas would use mail order due to the fact that they don't have access to commercial centers also the fact that it tends to be cheaper. If you are the type to have things delivered to you and not move around much you probably will be more open to the services of a mail order company or some kind of online shopping. If you are part of the younger generation in my opinion you would prefer to go to a shopping mall and browse around, or maybe you are a shopaholic and just buy things you don't need at the first sight of it, having access to a mail order company especially with the ease of making a purchase, you may find yourself leaving a significant dent in your wallet.

2.1.3 QUICK INSIGHTS

I had joined the datasets i.e. AZDIAS and Customers. I then added a label customer_identifer where all AZDIAS rows were labelled as 0 and Customer rows labelled as 1. I then used an XGBoost to model and predict if a row was a customer or not. I used shapely values to understand feature importance's as well as get a feel for what to expect when segmenting. This modelling perspective had also become a sort of an evaluation metric which I have used to judge the subsequent pre-processing steps taken by observing if the model became more or less predictive of customers using the pre-processed data.

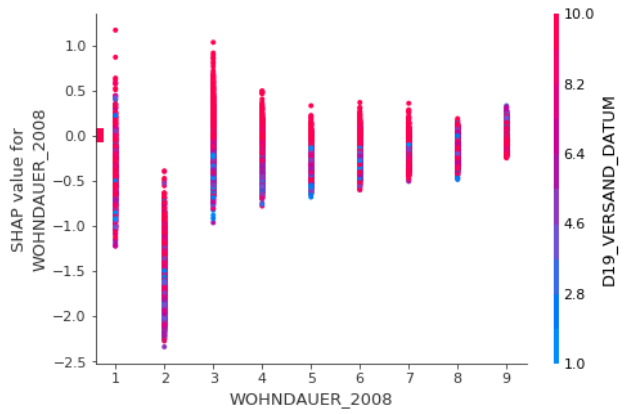

Figure 3: dependence plot of WOHNDAUER_2008

This gives us a clearer view of how having a length of residence below a year seems to result in a higher log odds of being a customer, but having a residency of 1-2 years results in a significantly lower log odds, and then an increase in log odds again in being a resident for more than 10 years. This could be because a year after people move, they start to try out other avenues of getting their products after which they return.

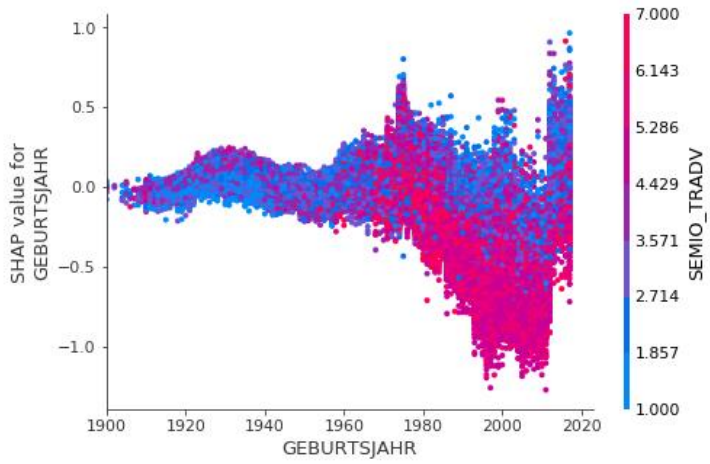

Figure 4: dependence plot of GEBURTSJAHR

This clearly shows younger people are less likely to be customers especially younger adults who have a very low affinity to being traditional minded. However, we see very young people have an increased odds of being customers, this is interesting, could it be because of inaccurate data? there seems to also be an area between 1920 and 1940 that has an increased probability of being a customer.

The interpretation of this would depend on what exactly does being traditional minded mean according to the data collectors.

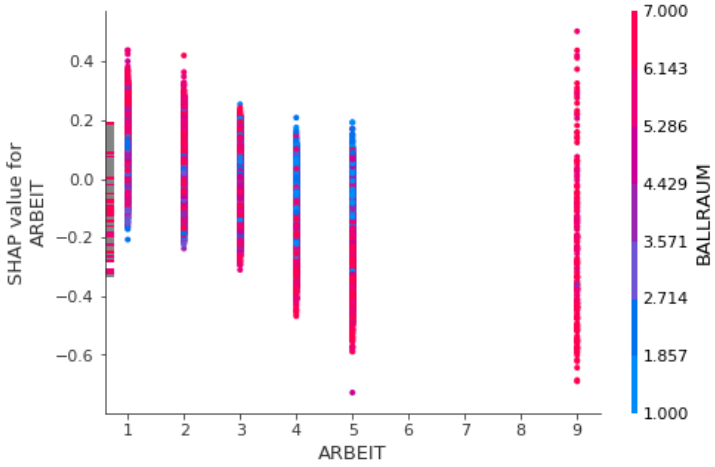

Figure 5: dependence plot of ARBEIT

We can see that if there is a larger share of unemployed people in the community then there likelihood of being a customer decreases, we also see in communities with larger number of unemployed living further away from urban centers decreases your odds of being a customer.

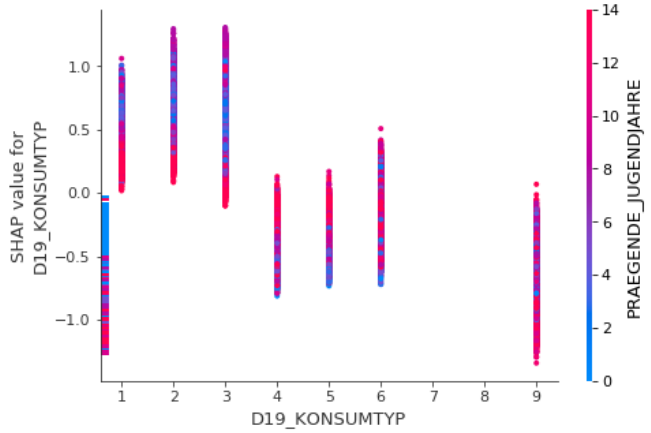

Figure 6: dependence plot of D19_KONSUMTYP

Being a universal, versatile and gourmet type consumer while being of an older generation seems to increase the chances of being a customer. While being family, informed or modern consumption type and part of the younger generation increases the probability of being a customer.

```
non_customer most frequent: 53
customer most frequent: 79
```

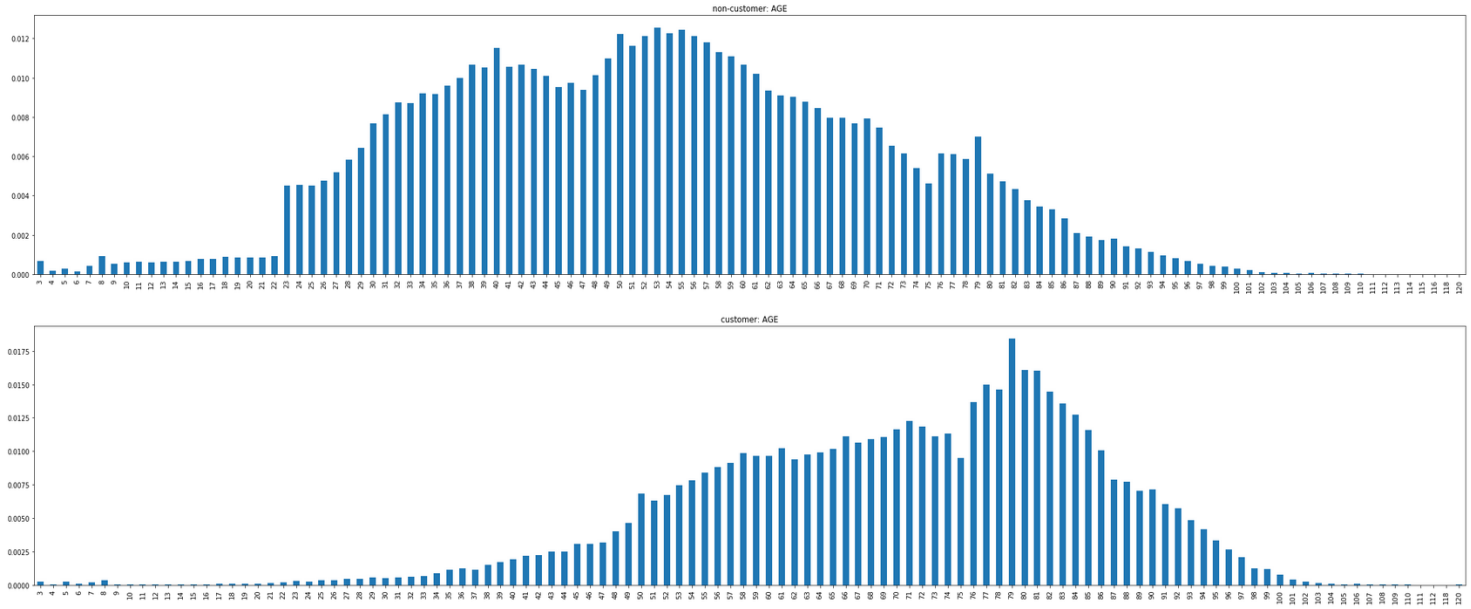

Figure 7: AGE distribution between customer and non-customer

In summary, customers are highly likely to be of an older age group who are traditional, male, are high income earners and are money wise. From the dependence plots above we see that being older doesn't necessarily increase your odds of being a customer but rather being younger decreases your odds of being a customer up to a certain point where the odds of being a customer increase significantly in seemingly very young individuals. This ties up in some way to our intuition developed above.

2.1.4 FEATURE ANALYSIS

During feature analysis it was found that 46 columns seemed to exist in the data dictionary yet were missing from the data frame it was later found that these columns had been incorrectly named in

Missing columns

Columns that exist in the data dictionary and not in the Dataframe

```
missing_cols = set(cleaned_mapping_df.Attribute) - set(azdias.columns)
print(missing_cols)
```

```
{'D19_KK_KUNDENTYP', 'PLZ', 'D19_SONSTIGE_RZ', 'D19_BANKEN_DIREKT_RZ', 'D19_HAUS_DEKO_RZ', 'D19_BEKLEIDUNG_REST_RZ', 'D19_BANKEN_GROSS_RZ', 'D19_LEBENSMITTEL_RZ', 'WACHSTUMSGEBIET_NB', 'GEOSCORE_KLS7', 'BIP_FLAG', 'D19_RAT
GEBER_RZ', 'D19_NAHRUNGSERGAENZUNG_RZ', 'D19_DROGERIEARTIKEL_RZ', 'D19_DIGIT_SERV_RZ', 'D19_SCHUHE_RZ', 'D19_VERSAND_REST_RZ', 'SOHO_FLAG', 'D19_KINDERARTIKEL_RZ', 'D19_LOTTO_RZ', 'D19_ENERGIE_RZ', 'D19_TIERARTIKEL_RZ', 'D
19_TELKO_MOBILE_RZ', 'D19_BILDUNG_RZ', 'D19_REISEN_RZ', 'D19_BIO_OEKO_RZ', 'D19_VERSICHERUNGEN_RZ', 'D19_WEIN_FEINKOST_RZ', 'CAMEO_DEUINTL_2015', 'D19_GARTEN_RZ', 'D19_TECHNIK_RZ', 'D19_BEKLEIDUNG_GEH_RZ', 'PLZ8', 'KBA13_C
CM_1400_2500', 'D19_SAMMELARTIKEL_RZ', 'HAUSHALTSSTRUKTUR', 'D19_TELKO_REST_RZ', 'D19_BANKEN_REST_RZ', 'GKZ', 'D19_KOSMETIK_RZ', 'D19_FREIZEIT_RZ', 'EINWOHNER', 'D19_VOLLSORTIMENT_RZ', 'D19_BUCH_RZ', 'D19_BANKEN_LOKAL_RZ',
'D19_HANDWERK_RZ'}
```

```
print(len(missing_cols))
```

```
46
```

Undocumented columns

columns that do not appear in the data dictionary , we dont have information for these columns

```
undoc_cols = set(azdias.columns) - set(cleaned_mapping_df.Attribute)
print(undoc_cols)
```

```
{'ALTER_KIND2', 'D19_TELKO_ONLINE_QUOTE_12', 'D19_KINDERARTIKEL', 'D19_KOSMETIK', 'STRUKTURTYP', 'D19_LOTTO', 'D19_SCHUHE', 'SOHO_KZ', 'KBA13_GBZ', 'EINGEFUEGT_AM', 'D19_TIERARTIKEL', 'EINGEZOGENAM_HH_JAHR', 'D19_REISEN',
'CJT_TYP_3', 'KK_KUNDENTYP', 'ALTERSKATEGORIE_FEIN', 'D19_SONSTIGE', 'D19_TELKO_REST', 'D19_BIO_OEKO', 'D19_BANKEN_LOKAL', 'D19_ENERGIE', 'D19_BANKEN_GROSS', 'DSL_FLAG', 'D19_BUCH_CD', 'MOBI_RASTER', 'RT_KEIN_ANREIZ', 'D19
_VERSAND_REST', 'HH_DELTA_FLAG', 'KBA13_BAUMAX', 'KBA13_KMH_210', 'VERDICHTUNGSRAUM', 'D19_BANKEN_REST', 'D19_VERSICHERUNGEN', 'D19_HANDWERK', 'D19_BANKEN_DIREKT', 'D19_BEKLEIDUNG_REST', 'KONSUMZELLE', 'D19_SOZIALES', 'VHA
', 'CJT_TYP_6', 'VK_DHT4A', 'D19_TECHNIK', 'FIRMENDICHTE', 'ALTER_KIND3', 'D19_FREIZEIT', 'VHN', 'ANZ_STATISTISCHE_HAUSHALTE', 'KBA13_CCM_1401_2500', 'D19_GARTEN', 'KBA13_ANTG1', 'D19_DIGIT_SERV', 'ANZ_KINDER', 'KBA13_HHZ
', 'AKT_DAT_KL', 'CJT_KATALOGNUTZER', 'CJT_TYP_5', 'KBA13_ANTG4', 'CJT_TYP_4', 'RT_UEBERGROESSE', 'UMFELD_JUNG', 'KBA13_ANTG3', 'VK_DISTANZ', 'CAMEO_INTL_2015', 'GEMEINDETYP', 'ALTER_KIND1', 'UMFELD_ALT', 'D19_WEIN_FEINKOS
T', 'D19_RATGEBER', 'ALTER_KIND4', 'KOMBIALTER', 'UNGLEICHENN_FLAG', 'D19_HAUS_DEKO', 'RT_SCHNAEPPCHEN', 'EXTSEL992', 'D19_KONSUMTYP_MAX', 'CJT_TYP_1', 'D19_SAMMELARTIKEL', 'customer_identifier', 'KBA13_ANTG2', 'D19_BILDUN
G', 'D19_BEKLEIDUNG_GEH', 'CJT_TYP_2', 'D19_VERSI_ONLINE_QUOTE_12', 'D19_LETZTER_KAUF_BRANCHE', 'D19_VOLLSORTIMENT', 'D19_NAHRUNGSERGAENZUNG', 'D19_LEBENSMITTEL', 'D19_TELKO_MOBILE', 'VK_ZG11', 'D19_DROGERIEARTIKEL'}
```

```
print(len(undoc_cols))
```

```
90
```

Figure 8: missing and undocumented features

the data frame and steps were taken to rename them.

This had left us with 314 columns documented out of 366.

A lookup function needed to be created to get summary statistics for a feature or set of features, the following is a sample output of the function.

```
**********************************************************************************************
AGER_TYP

Description:
 best-ager typology
Mapping:
 {-1: 'unknown', 0: 'no classification possible', 1: 'passive elderly', 2: 'cultural elderly', 3: 'experience-driven elderly'}
Additional Notes:
 ['in cooperation with Kantar TNS; the information basis is a consumer survey']
Info level:
 ['Person']

Column Summary
Value Counts Percentage
-1    71.071123
 0     1.197832
 1    11.098624
 2    13.329910
 3     3.302511
Name: AGER_TYP, dtype: float64

Unique Values:
5
**********************************************************************************************
AKT_DAT_KL

Feature info does not EXIST in data dictionary
Column Summary
Value Counts Percentage
1.0    47.077820
2.0     1.966990
3.0     2.529937
4.0     2.195733
5.0     3.075061
6.0     2.699301
7.0     2.093967
8.0     1.723563
9.0    25.547225
NaN    11.090405
Name: AKT_DAT_KL, dtype: float64

Unique Values:
10
```

Figure 9: output of lookup function

Features are grouped into information levels this is important as it gives a starting point from which to analyse features. These natural groupings allow inference to be made about some undocumented columns as well as gives us a better understanding of existing columns.

Since German is not a language that I speak some of the columns I have translated using google translate and thereafter used other correlated columns or columns in the same group to infer what those columns mean. This allowed me to document some undocumented columns further increasing the interpretability of the final results.

There are many missing values in the dataset where some columns have as much as 90 percent of their values missing. However, the values seem to not be missing at random (MAR) hence care has

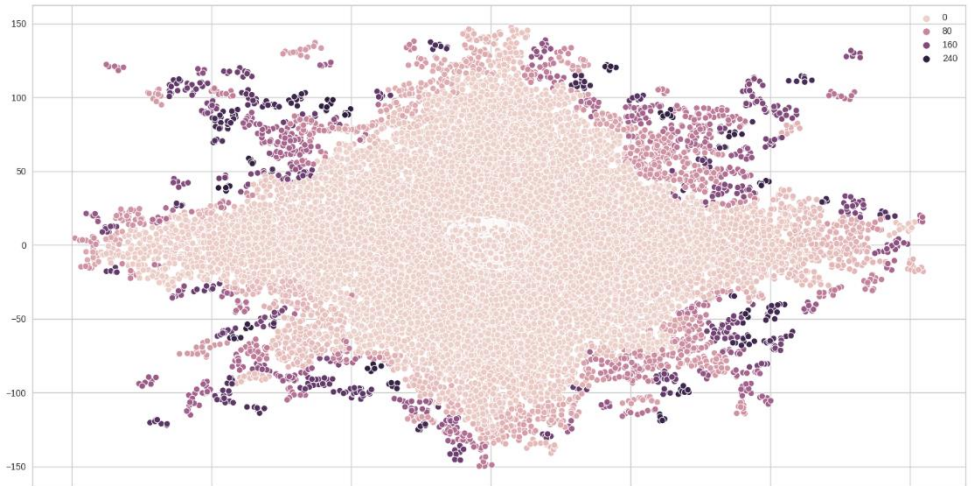

Figure 10: Using rapidsAI to do GPU accelerated dimensionality reduction and clustering via DBSCAN

to be taken when filling missing values as it might be that the missingness of the data carries some underlying information about the individual or data distribution.

2.2 ALGORITHMS AND TECHNIQUES

There are a few algorithms and methodologies I have utilized in this project I will discuss the key ones and their justifications here.

2.2.1 IMPUTATION

Imputation plays an important role in data analysis especially when there is a large amount of data missing in some sort of order. Simple imputation techniques may be sufficient but also risk introducing bias into the data. There are more advanced methods that seek to mitigate the bias and impute data in a more intelligent way utilizing the natural connectivity/correlation/dependence of data collected. I have chosen to come up with a custom imputation method using the XGBoost algorithm. Since it has the ability to handle missing data by selecting the node split that will minimize the loss, it is convenient in filling a single column using other columns regardless of missing values being present in the other columns.

Simple imputation techniques such as median and KNN was also tried, KNN due to the large number of features proves intractable for the memory available. A more advanced technique was also used viz. Iterative imputation , an implementation of MICE from the R package.

2.2.2 DIMENTIONALITY REDUCTION

The go to algorithm for dimensionality reduction is PCA, however from my understanding of PCA which seeks to attain a projected co-ordinate system that captures the most amount of variance using sum of squares, how would this affect categorical variables? After research I had found that there is a technique better suited to data with categorical variables which is Multiple correspondence analysis (MCA), a sort of PCA for categorical variables. This made sense since all the features in the data after pre-processing was categorical. A possible alternative to still use PCA is to one hot encode the entire dataset yet this is not the preferred method and does not intuitively make sense to me.

2.2.3 CLUSTERING

Same as above clustering also has a go to algorithm which is K-means, I would say probably because it is well documented and there exists many efficient implementations of it. I did find other algorithms such as K-modes with would have been ideal in this case but however the existing implementations are rather slow and inefficient. The reason this would have been ideal is that the K-modes algorithm uses the hamming distance to classify nearness of points which is a better distance metric to use for categorical data as opposed to Euclidean distance which K-means uses. The solution chosen was to use MCA and then K-means. DBSCAN was also considered due to some of its advantages over K-means such as its ability to seek out irregular shaped clusters in high dimensional space. However, this was discovered toward the end of this project and was not enough time to pursue this avenue and do it justice.

The elbow method was used to get the k value for the K-means algorithm.

2.2.4 CLASSIFICATION

The main models used during classification are:

- XGBoost
- CATBoost
- LightGBM
- Gradient Boosting Classifier

XGboost was used as a benchmark model in general as it has the ability to handle missing values allowing me to get evaluation results on raw unaltered data as well as pre-processed data.

PYCARET

Pycaret is a low code machine learning library that allows one to train and test a wide range of models with just a single line of code. The main reason this library was chosen was to get a quick over view of results of different models on the data in the least time possible. It also has a built-in grid search optimizer as well as a calibrator. Diagnostic diagrams can be produced easily, but I was only interested in its ability to generate a quick over view of the different models.

```
In [175]: compare_models(fold=5,sort="AUC")
```

Out[175]:

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa
0	Extreme Gradient Boosting	0.987600	0.770800	0.000000	0.000000	0.000000	0.000000
1	Gradient Boosting Classifier	0.987600	0.761000	0.000000	0.000000	0.000000	-0.000100
2	Ada Boost Classifier	0.987600	0.757600	0.000000	0.000000	0.000000	0.000000
3	Light Gradient Boosting Machine	0.987500	0.732800	0.000000	0.000000	0.000000	-0.000100
4	CatBoost Classifier	0.987600	0.721800	0.000000	0.000000	0.000000	0.000000
5	Naive Bayes	0.962800	0.706800	0.068200	0.118200	0.025300	0.015100
6	Quadratic Discriminant Analysis	0.929200	0.680800	0.115200	0.036600	0.044000	0.028500
7	Linear Discriminant Analysis	0.987600	0.678000	0.000000	0.000000	0.000000	0.000000
8	Extra Trees Classifier	0.987600	0.677800	0.000000	0.000000	0.000000	-0.000100
9	Logistic Regression	0.987600	0.674800	0.000000	0.000000	0.000000	0.000000
10	Random Forest Classifier	0.987600	0.565300	0.000000	0.000000	0.000000	0.000000
11	K Neighbors Classifier	0.987500	0.506300	0.000000	0.000000	0.000000	-0.000100
12	Decision Tree Classifier	0.972200	0.501900	0.032900	0.025900	0.028900	0.015100
13	SVM - Linear Kernel	0.984800	0.000000	0.009400	0.035800	0.014800	0.010100
14	Ridge Classifier	0.987600	0.000000	0.000000	0.000000	0.000000	0.000000

Figure 11: PYCARET

AUTO-SKLEARN

auto-sklearn is an automated machine learning toolkit and a drop-in replacement for a scikit-learn estimator. It enables training of multiple models and ensembles.

AUTO-GLUON

AutoGluon enables easy-to-use and easy-to-extend AutoML with a focus on deep learning and real-world applications spanning image, text, or tabular data. Intended for both ML beginners and experts, AutoGluon enables you to:

- Quickly prototype deep learning solutions for your data with few lines of code.
- Leverage automatic hyperparameter tuning, model selection / architecture search, and data processing.
- Automatically utilize state-of-the-art deep learning techniques without expert knowledge.
- Easily improve existing bespoke models and data pipelines, or customize AutoGluon for your use-case.

FAST.AI

Fast Ai is basically a framework for deep learning built on top of pytorch which makes proto-typing state of the art models with ease.

VTREAT

A 'data frame' processor/conditioner that prepares real-world data for predictive modelling in a statistically sound manner. 'vtreat' prepares variables so that data has fewer exceptional cases, making it easier to safely use models in production.

OPTUNA

Optuna, a Bayesian hyper parameter optimizer library was used to optimize stand-alone models. 5-fold stratified cross validation was used to evaluate the model's robustness and consistency.

2.3 BENCHMARK MODEL

The benchmark for the customer segmentation part of the problem is formulated from a supervised learning perspective, to facilitate a quick method of getting some kind of benchmark I fused the customer and population datasets labelling them respectively. After dropping columns with more than 80% missing values, I fitted an XGBoost model which achieved a 0.93 AUC score to predict the customer class. I then used Shapely values to get some quick insights into important features and their interactions.

For the mail order response prediction again an XGBoost model was fitted on the raw data with minor clean-up which attained an AUC score of 0.72 with the scale_pos_weight parameter set to cater for the class imbalance.

3. METHODOLOGY

3.1 DATAPREPROCESSING

3.1.1 DROPPED FEATURES

Many features have been dropped, out of the 366 features to begin with I am left with 101 features after feature engineering.

Features with very high missing values such as the ALTER_KIND columns were dropped, however the information for these columns are present implicitly in ANZ_KINDER. I created a column with the number of missing values from the 4 ALTER_KIND columns and checked the correlation with ANZ_KINDER and found a -0.97 correlation showing that the ALTER_KIND columns did not contain missing or unknown values but rather values signifying a lack of kids. I did check for rows where the number of kids was 1 or more with all missing values in ALTER_KIND, probably meaning parents did not disclose their kids age or some other reason.

Features grouped as PLZ8 seemed to have many correlated columns and seemed to explain already what was in the other groupings so was dropped.

Features in microcell RR3 was also dropped due to the noisiness and over powering of the components in the dimensionality reduction algorithm.

Columns that were undocumented and which I could not infer or translate was dropped.

Rows with missing values were not dropped as I believe that the “unknowness” of values carries some information about that datapoint.

High cardinality columns such as 'EINGEFUEGT_AM','ALTERSKATEGORIE_FEIN','MIN_GEBAEUDEJAHR','EINGEZOGENAM_HH_JAHR' were dropped. Some of them was dropped due to more complete and similar information existing in other columns.

Many household transaction columns were also dropped due to high correlation with each other, I could have in this instance used predictive power score to gauge which of the correlated features are more predictive of the target column and kept that one. Dropping these columns reduced model complexity as well as getting rid of some features I could not explain.

3.1.2 MISSING VALUES

Values that have mappings for unknown from within the attribute values datafile were mapped to unknowns. Except for certain cases where we could more intelligently infer the values.

D19_LETZTER_KAUF_BRANCHE has a category for unknown transactions. KONSUMNAEHE has a value that states the cell is a consumption cell so in rows where KONSUMNAEHE = 1 we set KONSUMZELLE = 1.

The rest of the columns were filled using the XGB imputer referenced earlier.

3.1.3 FEATURE ENGINEERING

GEBURTSJAHR was converted from birth year to current age, PRAEGENDE_JUGENDJAHRE was used to extract the generation the individual belongs to. The effect of the feature engineering was judged by looking at how well the new feature differentiates between customer and population.

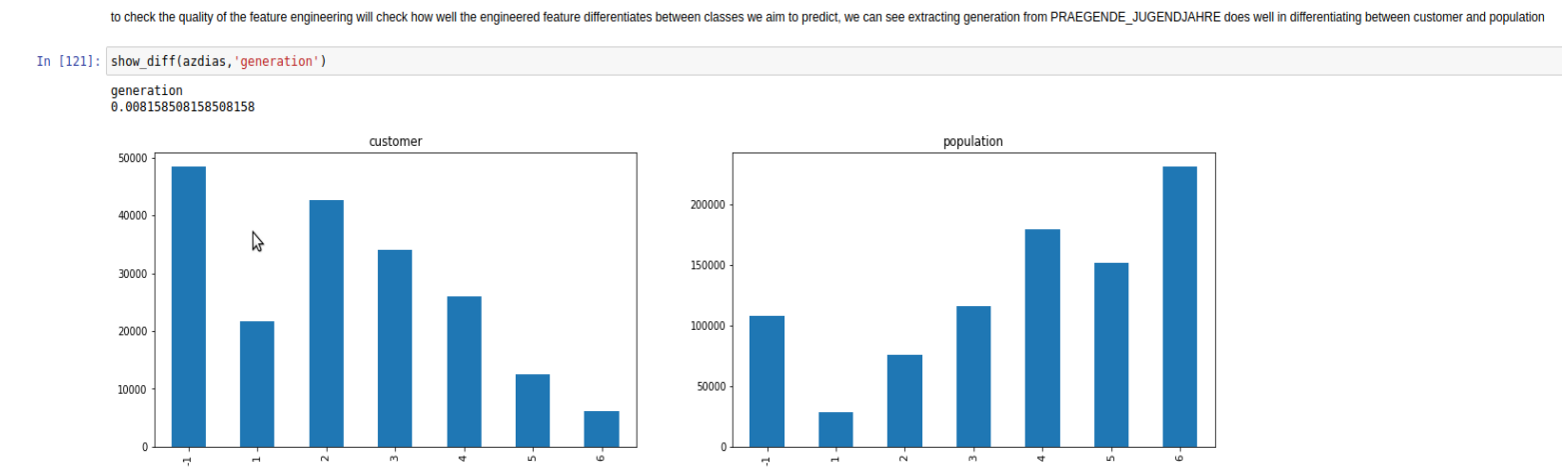

Figure 12: checking engineered features

I noticed many age columns so decided to do an integrity check to see if it all adds up, what I found was interesting.

Grouping the DF by generation and displaying the ages in each generation it was found that the generation was incorrectly classifying the age group e.g. generation 70ies described people from the ages of 65-55 who were born in the years 1964-1955. A remapping was done to correctly describe the generation values.

Like that other features were engineered and compared. Some features that seemed to be well suited as being flags such as FINANZ_TYP was retyped to category. One hot encoding of all features was avoided due to the fact that I planned on initially using the XGBoost model to predict and then explain the customers. Sparse data does not do well with tree-based algorithms. [https://towardsdatascience.com/one-hot-encoding-is-making-your-tree-based-ensembles-worse-heres-why-d64b282b5769]

Many of the numerical columns were binned into quantiles were possible and where not, into logical levels.

3.2 IMPLEMENTATION

3.2.1 K-MEANS

Before clustering dimensionality reduction is done, the MCA algorithm did not need the data to be pre-processed as it internally one hot encodes the data and carries out correspondence analysis on the indicator array. The package used to do MCA is not actively maintained so a few parts of the package had to be tweaked in order to be compatible with the latest pandas version.

```
In [37]: # Kmeans with PCA
         get_kmeans_elbow(np.array(azdias_mca),k=(5,20))
```

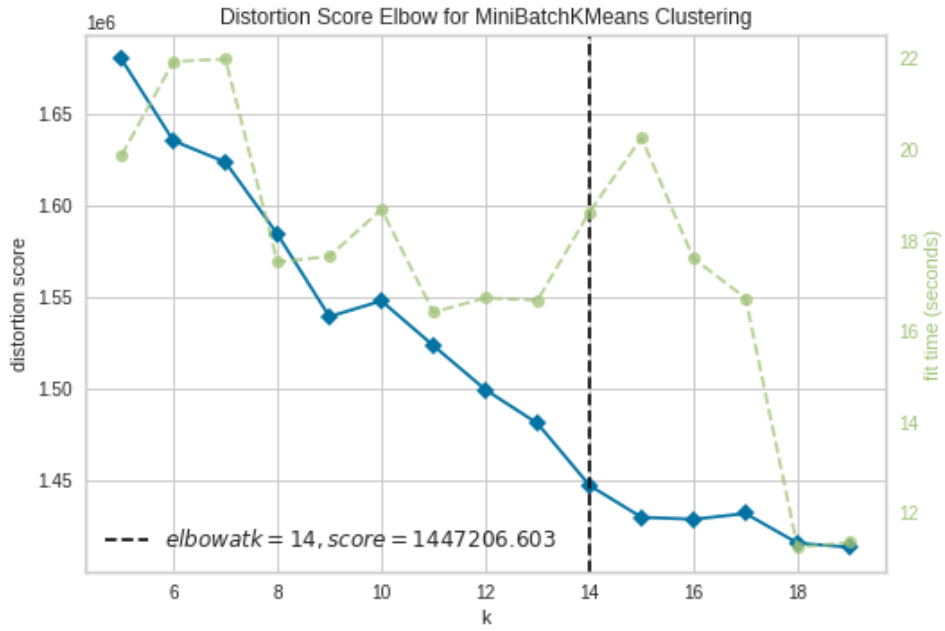

Figure 13: Kmeans Elbow Visualizer

The yellowbricks package which is part of the sklearn contrib provides a neat kElbowVisualizer which programmatically detects the and displays the Elbow in the elbow method used to select k clusters for the K-means algorithm.

3.2.2 CLASSIFICATION

For classification I have attempted to keep as many columns as possible as I find it gave a better score on the leader board, hover this goes against my philosophy of "explainability may outperform predictability in the long run" example, when the underlying distribution of our data changes while the high predicative model is deployed in production. How will one diagnose what has changed, draw insights and improve current business processes?

Auto-ML

For the auto ML libraries used little to no data preparation was needed as the library handles it for you allowing quick proto-typing and fairly good results.

Standard procedures

Columns defined as object were filled and converted to fully numerical columns, columns that were found to be inconsistent and/or high cardinality were dropped. Numeric columns were binned and features engineered from the first part was engineered here.

Highly correlated columns were dropped based on their predictive power score with the target variable. Duplicates were dropped from the majority class in the hope that the model generalizes better. XGBoost was used in all steps as a benchmark to judge if the predictability was getting better or worse.

Different sampling methods were tried as well as Bayesian optimization.

4. RESULTS

4.1 CUSTOMER SEGMENTATION

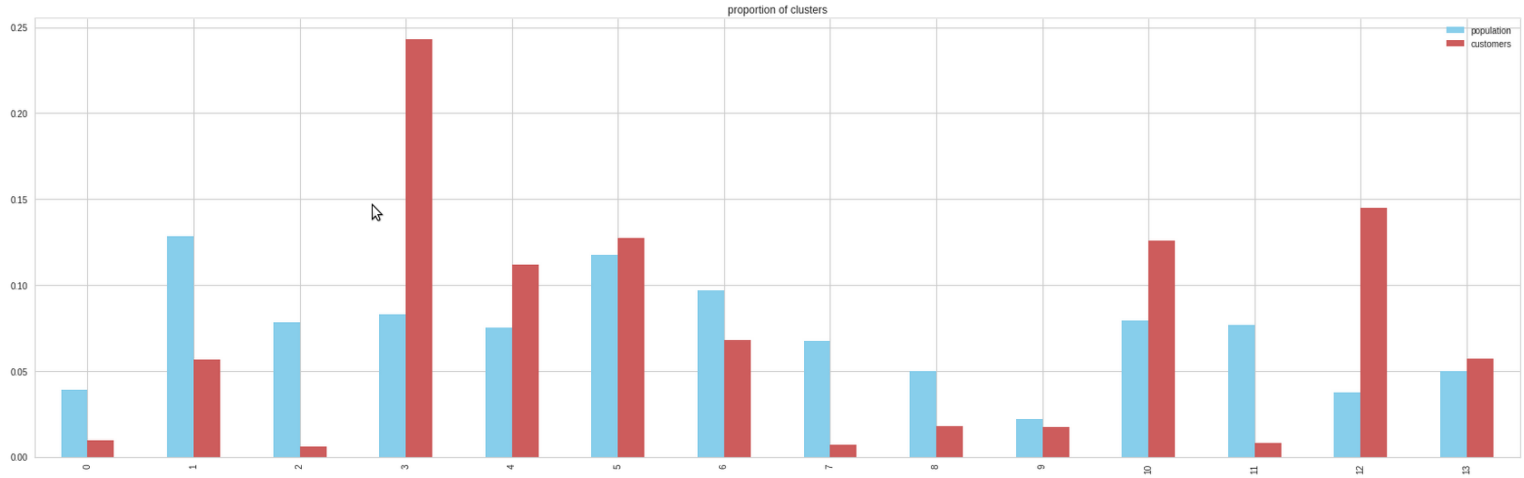

Figure 15: customers and population clusters

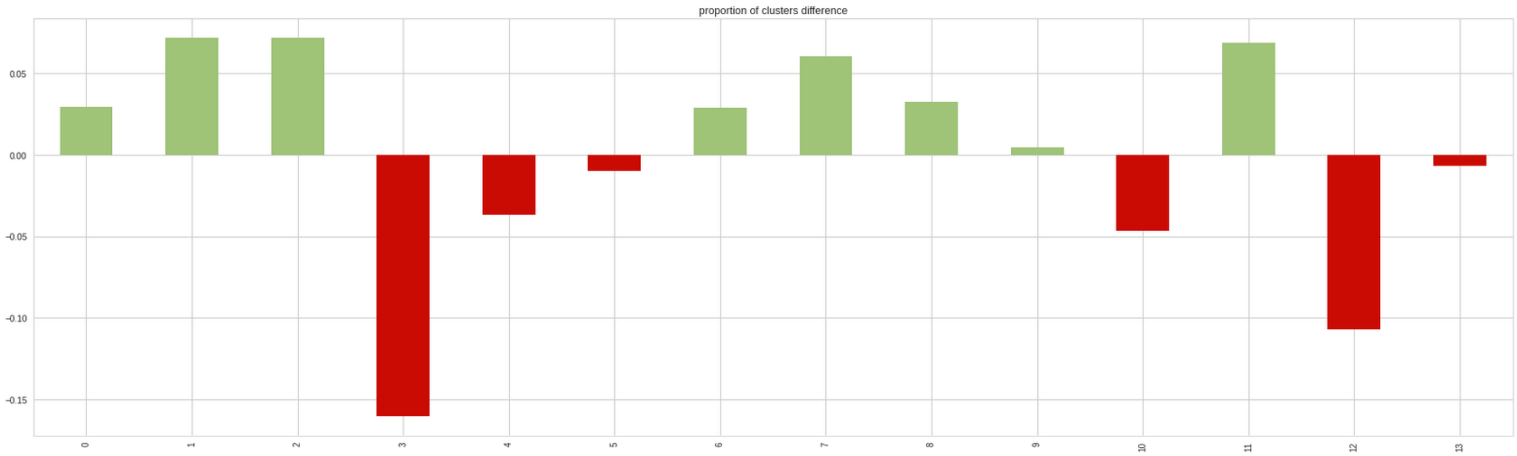

Figure 14: difference in proportion in each cluster

CUSTOMERS

LARGEST CUSTOMER GROUP

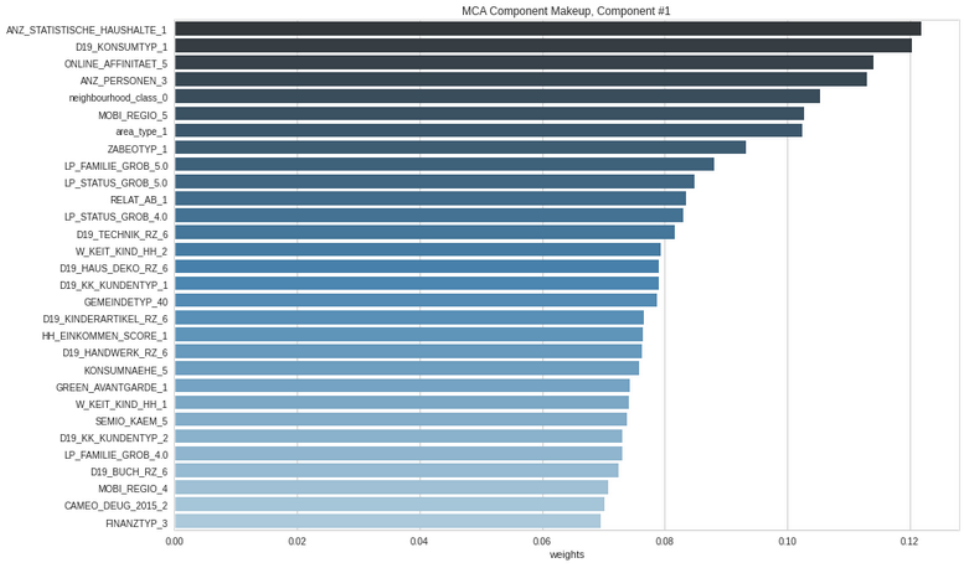

Figure 16:MCA component make up for largest customer group

The largest customer group consists of buildings with a single household i.e. people to probably have their own homes, primarily universal consumption type with very high online affinity. They tend to come from multi-person households or families. They have a very low mobility score this makes sense and ties up with the fact that they use the mail-order company. They live in wealthy neighbourhoods, and are generally conservationists. Interesting to note the municipal type is type 40, since we have no exact definition for this we would need to dig deeper, I did note that the lower values for municipal type seemed to be more urban areas, so based on that I will deduce that these people don't exactly stay in the city but rather a little bit away from the cities. They are in generally upper-class wealthy and family orientated individuals.

SECOND LARGEST CUSTOMER GROUP

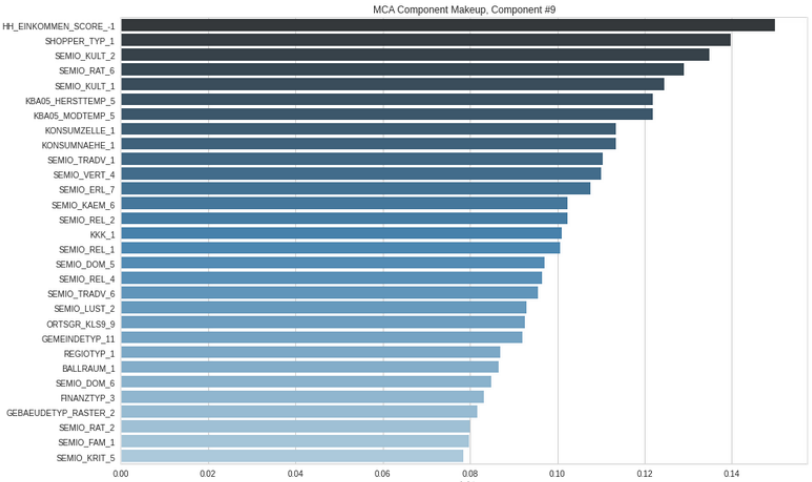

Figure 17: MCA component make up for second largest customer group

This group of people seem to be more reserved as they do not wish to divulge their household income information, they are culturally minded and traditional yet have a low affinity to rationale. They are located within consumption cells. They are religious and have a high purchasing power. There seems to be people who also have a low affinity to being traditional and a high affinity to being sensual, one could assume that these are younger people. These people are generally from municipal type 11 which according to the correlation with community size says that these people are from urban areas, closer to cities. Their Finanz type is "main focus own house" and are in mixed cells with high business share, this also affirms that these people are close to cities but not city dwellers as such. They come from upper class neighbourhoods. They are also classified as wealthy people but seemingly one wealth classification lower.

THIRD LARGEST CUSTOMER GROUP

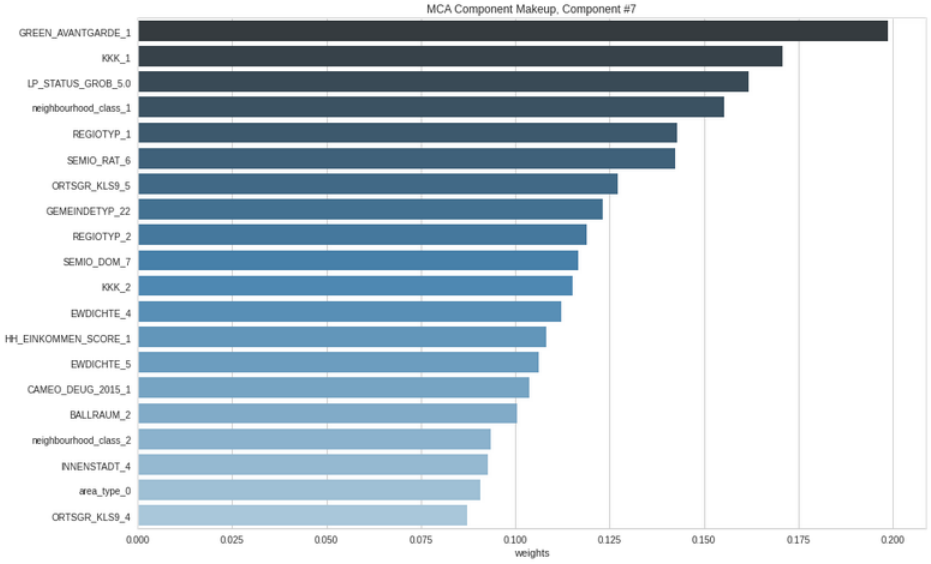

Figure 18: MCA component make up for third largest customer group

This group seems to be rural/outskirt dwellers who are indifferent to conservation, with high religious affinity. They also seem to be foreign to assimilated names and not German nationals, they have hedonistic tendencies and tend to be more demand shopping orientated. Possible these people are vagabonds living in industrial areas who can fend for themselves and live month to month or day to day.

NON-CUSTOMERS

SMALLEST CUSTOMER GROUP

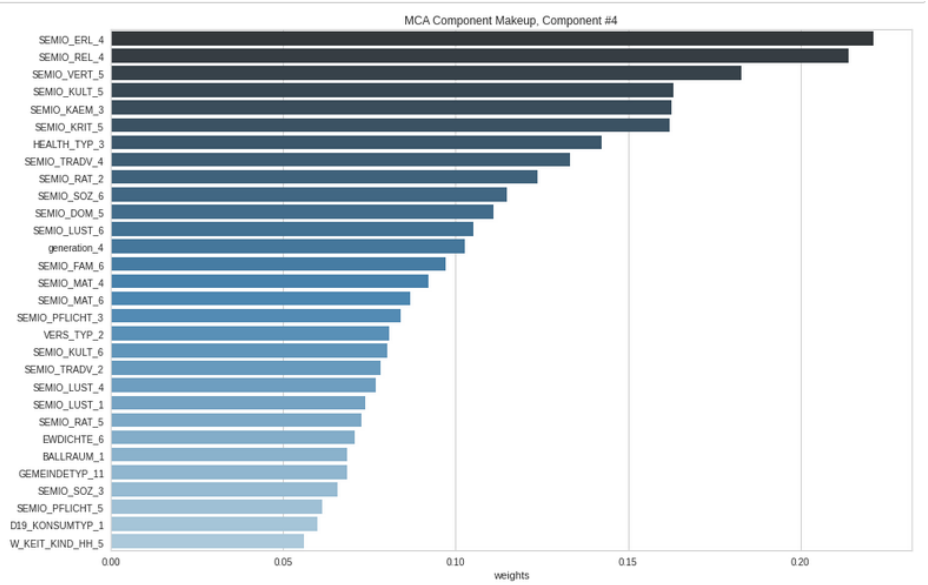

Figure 19: MCA component male up for smallest customer group

These people are in the age group of 65-55, are described as being fightful, jaunty hedonists and rationally minded people. They tend have a low affinity to social mindedness. They are unlikely to have a child present in their household.

SECOND SMALLEST CUSTOMER GROUP

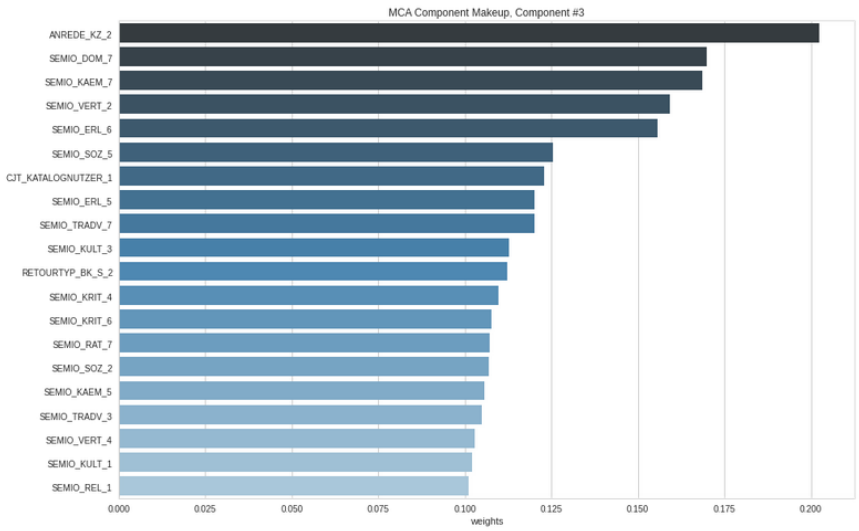

Figure 20: MCA component make up for second smallest customer group

This group is described as being females. They rarely use catalogues and have low affinity to tradition. They are demanding heavy returners, probably why they don't use the mail order company or is this trait just a consequence of femininity, a generalisation. they have low affinity to being critical and rational minded.

The above analysis seems to be sensible in that it ties in well with the intuition developed at the beginning of the project.

A secondary method was used to evaluate customers the reader may refer to the end of the notebook if interested. Using the method of formulating the problem from a supervised setting, a model was trained to predict accurately the customer class then shapely values were used to explain what the model uncovered.

4.2 CLASSIFICATION

After having trained and tested a variety of models and methods including auto-ml methods one striking conclusion I have made is that there is much value in good feature engineering. However, feature engineering takes a lot of time and effort from the data scientist, whereas autoML methods generally give good enough results with very little effort.

My objective when it came to this part of the competition was primarily to get exposure to as many as possible methods and techniques used to model and predict on a real life dataset, often I would fine myself discovering a new tool after which I would spend a few days trying it on a range of different problems. Having said that the most well performing model on the leader board and the model that I found generalized the best in general was the CATBOOST classifier combined with under sampling methods.

The XGB models used for benchmarking usually scored in the range of 0.72-0.75 AUC score on the hold out test set. The CATBoost model scores in the range of 0.77 to 0.80 AUC score on the test set. All models trained with the scale_pos_weight parameter set to the recommended setting of (sum of negative class)/(sum of positive class) showed better generalization to the test set even if they scored a lower validation score on the validation set.

5. CONCLUSION

I can confidently say after having spent the time and effort on this project I have definitely become much more confident and knowledgeable in the field of machine learning and data science. Being the first time working on a real-life data set and having to be involved in the data cleaning process and data reasoning process was of great value to me.

This project was the challenging project I needed to really propel my understanding of the ML workflow. It also exposed me to big data tools like DASK and Vaex but optimizing the dataset and working with pandas eventually the easier option. I faced and discovered many “gotchas” that I had always read about but no had the chance to experience and combat them first hand.

Some improvements I would definitely consider is documenting a more structured way the classification task with the models and methods used, I guess having a global look at what works and what doesn’t enables one to incrementally improve on model accuracy or AUC in our case. For the segmentation task, having the skills I have developed over the course of this project, I would love to be able to go back and explore the data again. I feel that there is still much insights to be extracted from the data that can facilitate more intelligent feature engineering and really provide an edge to the machine learning tasks.

I have learned more in this course then I have learned before and would definitely recommend this to any aspiring machine learning engineers or data scientists.