ARVATO CUSTOMER SEGMENTATION REPORT

Contents

ARVATO CUSTOMER SEGMENTATION REPORT

Project Overview

This project is a data science task assigned by Arvato Financial Solutions that involves analyzing demographics data for a mail order company's customer base and the general population to identify parts of the population that describe the core customer base. There are four parts of the project:

- Data Preprocessing
  Data provided by Arvato comes in a raw format which needs to be preprocessed to make it viable for unsupervised and supervised learning tasks in the later sections. Data is first explored get information on feature types, feature correlations, missing data, formatting issues etc. Visualizations and descriptive statistics are generated in this part. This information is then applied to build a singular data cleaning function to clean all datasets. Data is then transformed into a form that can be easily fed into machine learning pipelines. Transformations are applied individually to numeric, ordinal and nominal data types.
- Customer Segmentation
  In this section, we apply principal component analysis to decompose data in order to reduce dimensionality. And finally, we apply K-means clustering to separate customers from the general population. These clusters are then analyzed to inspect how features differ between customers and the general population
- Predictive Modeling
  This section involves building a classifier that predicts if an individual in the general population will become a customer. The classifier is parameterized using grid search which uses internal cross validation with an error measure of 'roc_auc' to acquire the best parameters for the classifier. A ROC curve is generated using prediction probabilities generated during cross validation.
- Kaggle Competition
  The model is applied to MAILOUT_TEST and the resulting probabilities are uploaded to the Kaggle competition to check performance.

Problem Statement

A mail-order company in Germany is looking into expanding their customer base. The best way to acquire new customers is by launching a targeted marketing campaign wherein specific individuals who are more likely to become customers are targeted with advertisements. The company requires identification of these 'potential customers' within the general population. Arvato Financial Solutions has tasked me with analyzing demographics data for both the mail-order company's customer base and the general population at large to identify parts of the population that best describe the core customer base of the company. The end goal of this project is to use data analysis from EDA and customer segmentation to build a model that

predicts whether an individual in the general population is likely to become a customer. There are two main objectives of this project:

- Perform customer segmentation in order to identify parts of population that best describe core customer base of the company.
- Use analysis from customer segmentation to develop a predictive model to identify which individuals in the general population are potential customers.

Metrics

I will use Area Under Curve (AUC) of Receive Operating Characteristic (ROC) curve as the metric for evaluation. The ROC curve plots the true positive rate (tpr) against the false positive rate (fpr). There is an inherent trade-off between the tpr and fpr: the higher the recall (tpr), the more false positives the classifier produces. A good classifier will capture as many of the actual customers as possible very early in the ROC curve sweep (the curve will be pushed against the top left corner). This metric will be used to parameterize the classifier during grid search in order to produce a classifier that has a ROC-AUC score as close to 1.0 as possible.

Data Exploration

The following figures show data-frame samples for all 4 datasets.

	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	ALTER_KIND1	ALTER_KIND2	ALTER_KIND3	ALTER_KIND4
0	910215	-1	NaN	NaN	NaN	NaN	NaN	NaN
1	910220	-1	9.0	0.0	NaN	NaN	NaN	NaN
2	910225	-1	9.0	17.0	NaN	NaN	NaN	NaN
3	910226	2	1.0	13.0	NaN	NaN	NaN	NaN
4	910241	-1	1.0	20.0	NaN	NaN	NaN	NaN

Azdias sample

	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	ALTER_KIND1	ALTER_KIND2	ALTER_KIND3	ALTER_KIND4
0	9626	2	1.0	10.0	NaN	NaN	NaN	NaN
1	9628	-1	9.0	11.0	NaN	NaN	NaN	NaN
2	143872	-1	1.0	6.0	NaN	NaN	NaN	NaN
3	143873	1	1.0	8.0	NaN	NaN	NaN	NaN
4	143874	-1	1.0	20.0	NaN	NaN	NaN	NaN

Customers Sample

ARVATO CUSTOMER SEGMENTATION REPORT

	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	ALTER_KIND1	ALTER_KIND2	ALTER_KIND3	ALTER_KIND4
0	1763	2	1.0	8.0	NaN	NaN	NaN	NaN
1	1771	1	4.0	13.0	NaN	NaN	NaN	NaN
2	1776	1	1.0	9.0	NaN	NaN	NaN	NaN
3	1460	2	1.0	6.0	NaN	NaN	NaN	NaN
4	1783	2	1.0	9.0	NaN	NaN	NaN	NaN

MAILOUT_TRAIN sample

	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	ALTER_KIND1	ALTER_KIND2	ALTER_KIND3	ALTER_KIND4
0	1754	2	1.0	7.0	NaN	NaN	NaN	NaN
1	1770	-1	1.0	0.0	NaN	NaN	NaN	NaN
2	1465	2	9.0	16.0	NaN	NaN	NaN	NaN
3	1470	-1	7.0	0.0	NaN	NaN	NaN	NaN
4	1478	1	1.0	21.0	NaN	NaN	NaN	NaN

MAILOUT_TEST sample

All 4 datasets contain the same 365 demographic features. All datasets also contain an individual identifier column named 'LNR'. Compared to AZDIAS, CUSTOMERS contains 3 additional non-demographic features named CUSTOMER_TYPE, PURCHASE_TYPE and ONLINE_PURCHASE. Compared to AZDIAS, MAILOUT_TRAIN contains 1 extra Boolean feature named 'RESPONSE'.

Features have to be manually assigned the correct data types. Additionally, most features have null placeholders e.g. values that are placeholders for NaN. Data types and these null placeholders have to be defined for each feature. The figure below shows results of manual inspection. Data types and null placeholders are now associated with feature names

feature	type	nan_indicators	action
AGER_TYP	nominal	[-1.0]	
AKT_DAT_KL	ordinal		
ALTER_HH	ordinal	[0.0]	
ALTER_KIND1	numeric		corr
ALTER_KIND2	numeric		corr
ALTER_KIND3	numeric		corr
ALTER_KIND4	numeric		corr
ALTERSKATEGORIE_FEIN	ordinal	[0.0]	
ANZ_HAUSHALTE_AKTIV	numeric		corr
ANZ_HH_TITEL	numeric		
ANZ_KINDER	numeric		
ANZ_PERSONEN	numeric		
ANZ_STATISTISCHE_HAUSHALTE	numeric		corr
ANZ_TITEL	numeric		
ARBEIT	ordinal	[9.0]	
BALLRAUM	ordinal	[-1.0]	
CAMEO_DEU_2015	nominal		format
CAMEO_DEUG_2015	nominal	[-1.0]	format
CAMEO_INTL_2015	ordinal		format

Data Types and null placeholders

ARVATO CUSTOMER SEGMENTATION REPORT

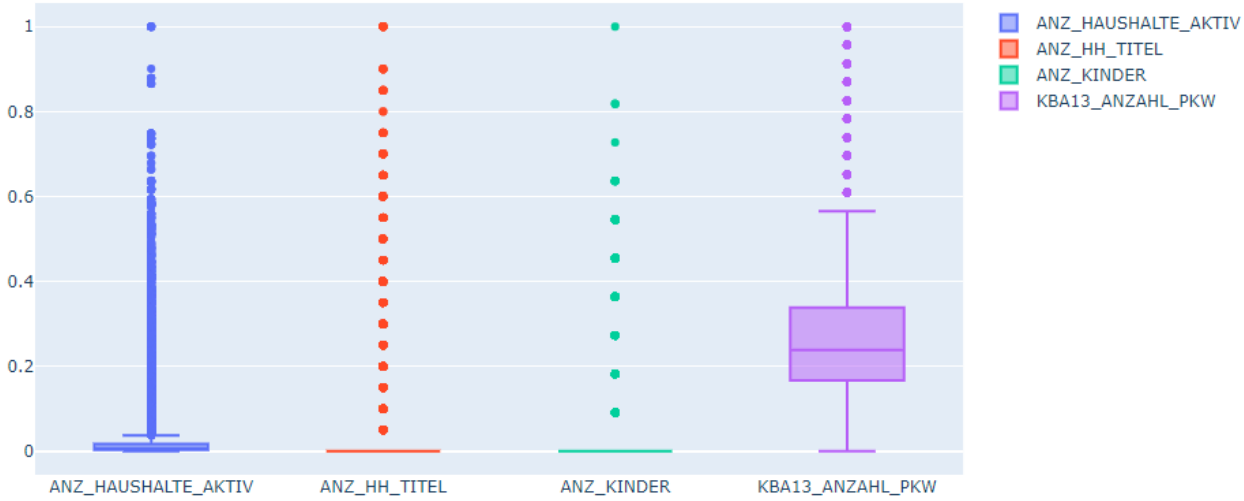
Visualizations

Statistical summaries of numerical features are shown in the table below. ANZ_HAUSHALTE_AKTIV has a mean of 8.287 and a median of 4. This means that 50% of the data has a value of or less than 4 while the average value lies further to the right possibly due to outliers. Distribution is right skewed. ANZ_HH_TITEL and ANZ_KINDER are very similar in that the distribution is right skewed. For these two features, it also seems that the distribution pivots at around 0 with outliers to the right.

	ANZ_HAUSHALTE_AKTIV	ANZ_HH_TITEL	ANZ_KINDER	ANZ_PERSONEN	ANZ_TITEL	KBA13_ANZAHL_PKW	VERDICHTUNGSRAUM
count	798073.000000	794213.000000	817722.000000	817722.000000	817722.000000	785421.000000	793947.00000
mean	8.287263	0.040647	0.154018	1.727637	0.004162	619.701439	4.58576
std	15.628087	0.324028	0.502389	1.155849	0.068855	340.034318	8.47152
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000
25%	1.000000	0.000000	0.000000	1.000000	0.000000	384.000000	0.00000
50%	4.000000	0.000000	0.000000	1.000000	0.000000	549.000000	1.00000
75%	9.000000	0.000000	0.000000	2.000000	0.000000	778.000000	5.00000
max	595.000000	23.000000	11.000000	45.000000	6.000000	2300.000000	45.00000

Numeric Feature Correlations

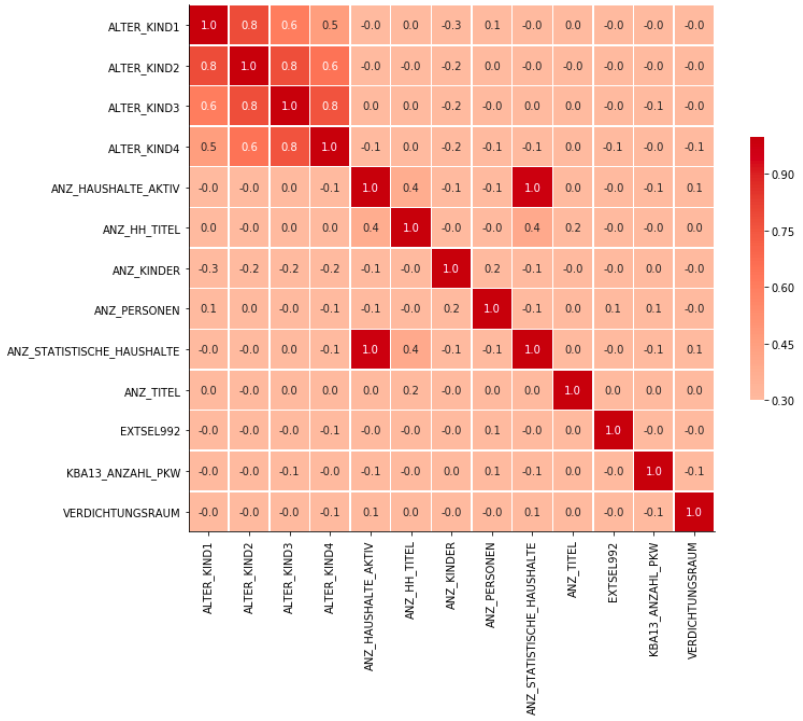
Min-max scaling is applied to the three features and box and whiskers plots are generated to visualize spread. KBA13_ANZAHL_PKW is also shown for contrast. ANZ_HAUSHALTE_AKTIV and ANZ_HH_TITEL are severely right skewed (skew > 10) as there are many strong outliers. ANZ_KINDER is pivoted at 0 but is less skewed due to a lower number of outliers. KBA13_ANZAHL_PKW is only slightly skewed to the right. There are some outliers to the right and the data is relatively well pivoted at the median. Overall, most numerical features are skewed.



Box and whiskers plot

ARVATO CUSTOMER SEGMENTATION REPORT

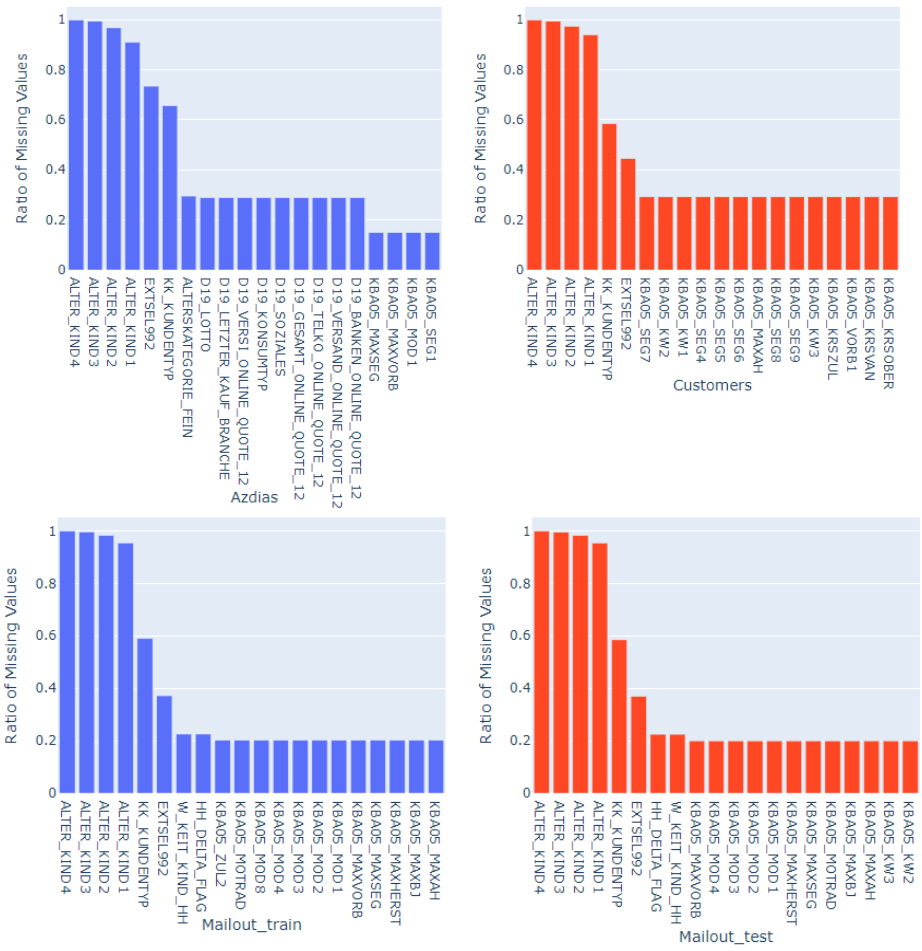
The heatmap below shows correlations between numerical features (Pearson correlation coefficient). There is strong correlation among the four ALTER_KIND features. ANZ_STATISTISCHE_HAUSHALTE and ANZ_HAUSHALTE_AKTIV also have a very strong correlation.



Heat Map

The following subplots show demographic features with the highest ratio of missing values in the 4 datasets. We can see that ALTER_KIND1, ALTER_KIND2, ALTER_KIND3, ALTER_KIND4, EXTSEL992 and KK_KUNDENTYP have >30 % of their data missing in all 4 datasets.

ARVATO CUSTOMER SEGMENTATION REPORT



Missing Values

Data Preprocessing

Data Cleaning

Using insights from exploratory data analysis and from summary statistics and visualizations, a data preprocessing function is developed to apply cleaning and formatting actions. The function applies the following actions in order:

I. Drop duplicated rows

II. Format corrupted columns

III. Reorder numerical encodings for disordered features

IV. Drop unrelated categories in ordinal features

V. Convert datetime feature values to standardized date categories

VI. Apply log function to skewed numerical features

ARVATO CUSTOMER SEGMENTATION REPORT

VII. Drop numerical features with very high correlations to other features

VIII. Replace null placeholders with NaN (not a number) in all features

IX. Drop features with > 30% of their data missing

X. Drop rows with >25% of their data missing (exclusively for CUSTOMERS and AZDIAS datasets).

The cleaning progress report is shown below. In addition to in place changes within the datasets, the same 17 columns are also dropped across all 4 datasets.

```
Cleaning azdias sample ...
Columns dropped:  15
['LNR', 'AGER_TYP', 'EXTSEL992', 'KK_KUNDENTYP', 'ALTER_KIND4', 'KBA05_BAUMAX', 'VHA', 'ANZ_STATISTISCHE_HAUSHALTE', 'TITEL_K
Z', 'ALTER_KIND3', 'GEBURTSJAHR', 'ALTER_KIND2', 'ALTER_KIND1', 'ALTER_HH', 'ALTERSKATEGORIE_FEIN']


Cleaning customers sample ...
Columns dropped:  17
['LNR', 'AGER_TYP', 'EXTSEL992', 'KK_KUNDENTYP', 'REGIOTYP', 'ALTER_KIND4', 'KBA05_BAUMAX', 'VHA', 'ANZ_STATISTISCHE_HAUSHALT
E', 'TITEL_KZ', 'ALTER_KIND3', 'KKK', 'GEBURTSJAHR', 'ALTER_KIND2', 'ALTER_KIND1', 'ALTER_HH', 'ALTERSKATEGORIE_FEIN']


Cleaning mailout_train sample ...
Columns dropped:  14
['LNR', 'AGER_TYP', 'EXTSEL992', 'KK_KUNDENTYP', 'ALTER_KIND4', 'KBA05_BAUMAX', 'VHA', 'ANZ_STATISTISCHE_HAUSHALTE', 'TITEL_K
Z', 'ALTER_KIND3', 'GEBURTSJAHR', 'ALTER_KIND2', 'ALTER_KIND1', 'ALTER_HH']


Cleaning mailout_test sample ...
Columns dropped:  14
['LNR', 'AGER_TYP', 'EXTSEL992', 'KK_KUNDENTYP', 'ALTER_KIND4', 'KBA05_BAUMAX', 'VHA', 'ANZ_STATISTISCHE_HAUSHALTE', 'TITEL_K
Z', 'ALTER_KIND3', 'GEBURTSJAHR', 'ALTER_KIND2', 'ALTER_KIND1', 'ALTER_HH']


ALL COLUMNS DROPPED:

['AGER_TYP' 'ALTERSKATEGORIE_FEIN' 'ALTER_HH' 'ALTER_KIND1' 'ALTER_KIND2'
 'ALTER_KIND3' 'ALTER_KIND4' 'ANZ_STATISTISCHE_HAUSHALTE' 'EXTSEL992'
 'GEBURTSJAHR' 'KBA05_BAUMAX' 'KKK' 'KK_KUNDENTYP' 'LNR' 'REGIOTYP'
 'TITEL_KZ' 'VHA']



Additional feature(s) dropped from azdias:  ['KKK', 'REGIOTYP']
Additional feature(s) dropped from customers:  []
Additional feature(s) dropped from mailout_train:  ['ALTERSKATEGORIE_FEIN', 'KKK', 'REGIOTYP']
Additional feature(s) dropped from mailout_test:  ['ALTERSKATEGORIE_FEIN', 'KKK', 'REGIOTYP']
```

Data Preprocessing

ARVATO CUSTOMER SEGMENTATION REPORT

Data Transformation

In addition to these data cleaning functions, further transformations need to be applied to the data before it can be fed into machine learning pipelines. The transformation pipeline is shown below. Feature Union contains three transformers that apply transformations to each of the three data types. Numeric pipeline first selects the features that are numeric and then imputes missing values. Missing data in each column is replaced with the column median. We use median and not the mean because the data was highly skewed. Next, the data is scaled to unit variance and 0 mean. Ordinal pipeline selects the ordinal features and applies imputation. Iterative imputer is used instead of a simple imputer because it involves an iterative process of probabilistically estimating missing values based on observed information from across your data set. The data is ordinally encoded next. Nominal pipeline selects nominal features and imputes missing values with the most frequent values in the respective columns. This is the best imputation method since values are not in a continuous range so iterative imputing cannot be applied. Finally, data is one hot encoded (binary encoding).

```
pipe_transform = Pipeline([

    ('features', FeatureUnion([

        ('numeric_pipeline', Pipeline([
            ('selector', AttributeSelector(numeric_features)),
            ('imputer', SimpleImputer(strategy='median')),
            ('std_scalar', StandardScaler())

        ])),

        ('ordinal_pipeline', Pipeline([
            ('selector', AttributeSelector(ordinal_features)),
            ('imputer', CustomImputer()),
            ('encoder', OrdinalEncoder(categories=ole_cat))

        ])),

        ('nominal_pipeline', Pipeline([
            ('selector', AttributeSelector(nominal_features)),
            ('imputer', SimpleImputer(strategy='most_frequent')),
            ('encoder', OneHotEncoder(categories = ohe_cat, sparse=False, n_values='auto'))
        ]))

    ]))

])
```

Pipeline

ARVATO CUSTOMER SEGMENTATION REPORT

Implementation

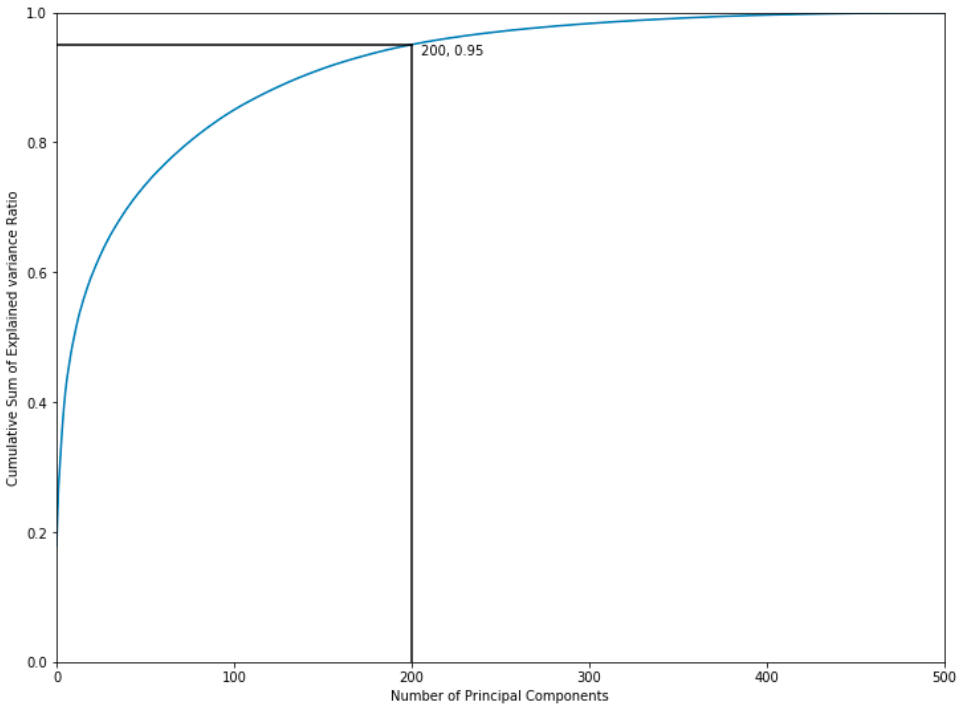
Customer Segmentation

Clean and transformed demographic data for is now available. The goal of this section is to identify parts of the population that best describe core customer base of the company. First, principal component analysis is applied to reduce dimensionality. Then, K-Means clustering is applied to separate customers and the general population into different clusters. This will allow for identification of features that are important in gauging customer behavior.

Principal Component Analysis

PCA is used to decompose a dataset with many (presumably) linearly dependent features into a set of linearly uncorrelated variables called principal components. With PCA, a complex dataset with many features can be represented in a much lower feature space. This transformation is defined in such a way that the first principal component accounts for most of original dataset's variability (it has the largest possible variance), and each succeeding component in turn has the highest variance with the condition that it be orthogonal to all preceding components.

The following plot shows how much of the original dataset's variability is retained at different numbers of principal components. The first 200 principal components capture 95% of the total variance in the original dataset so we select 200 principal components moving forward.



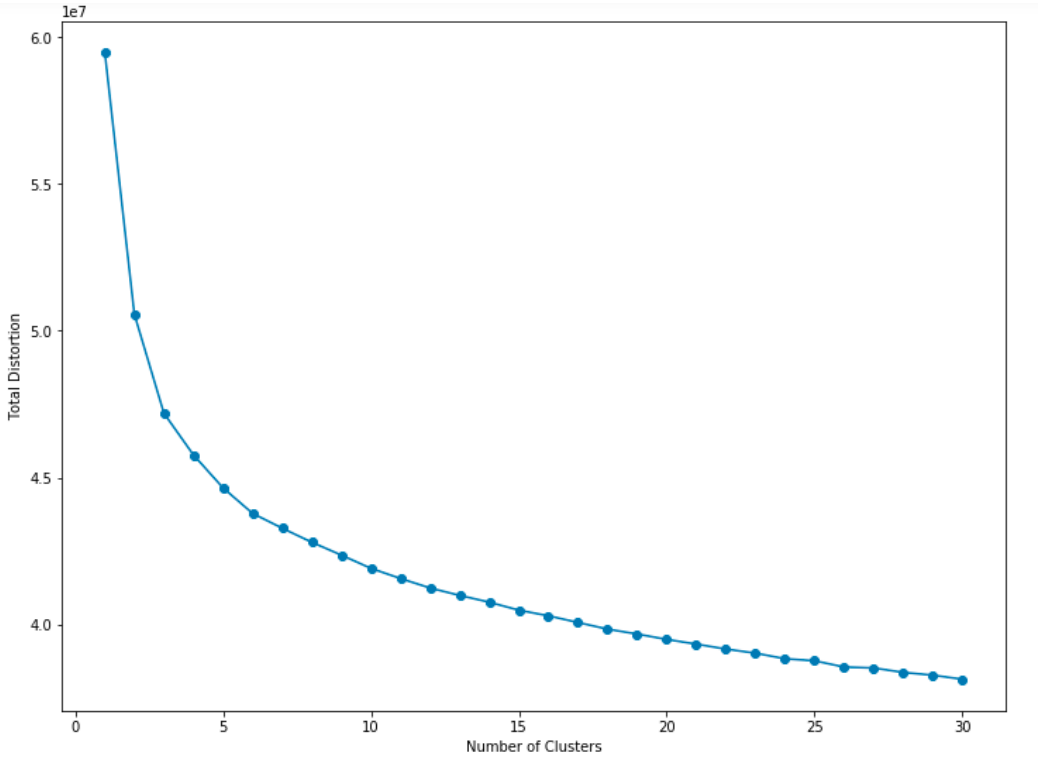
PCA

ARVATO CUSTOMER SEGMENTATION REPORT

K-Means Clustering

K-Means clustering is an unsupervised learning algorithm that partitions a data into a defined number of clusters. First, the algorithm identifies a number of centroids and then assigns each observation to the closest centroid. The algorithm tries to minimize the within-cluster sum of squares (distortion). The process is repeated over many iterations until the centroids have stabilized.

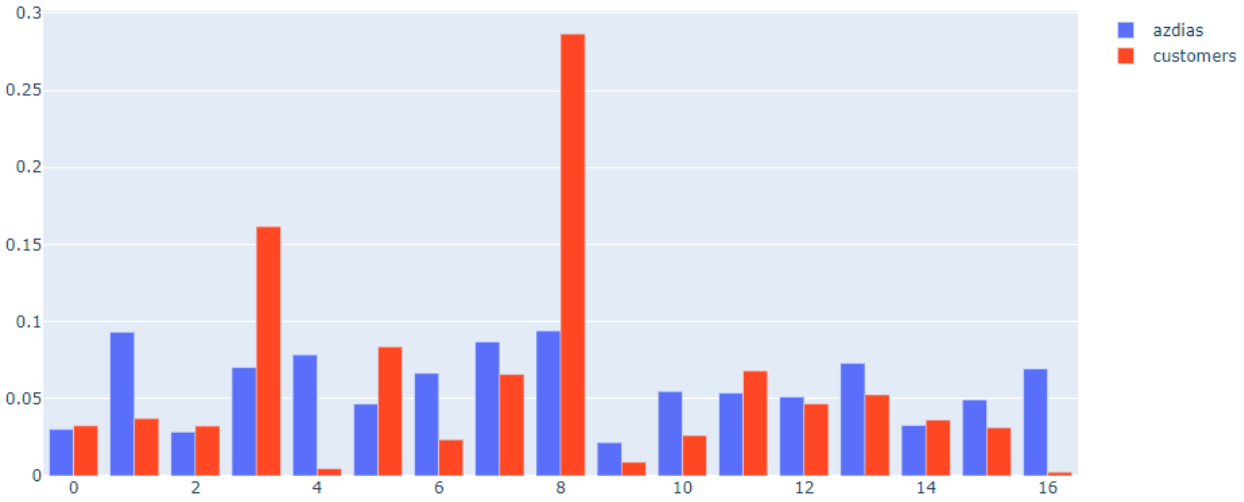
K-Means clustering is applied to the decomposed AZDIAS dataset. The plot below shows total distortion at different numbers of clusters. As we increase the number of clusters, distortion decreases. Notice that the graph appears to be heading towards a local minimum around N=14 before it breaks this pattern. The same pattern is located around N=17 although it is harder to notice. Repeated testing with random samples of the datasets show that the graph becomes almost linear after N=17 and the rate at which distortion decreases becomes small. N=17 clusters is the optimal number of clusters for our K-Means algorithm.



K-Means Clustering

The plot below shows the proportion of data contained in each cluster. We can see that the clustering has worked well since more than 45% of all customers are contained in two clusters only: clusters 3 and 8. Customers are overrepresented in these clusters as compared to the general population. Conversely, customers are very underrepresented in cluster 4 and 16. We can find cluster centers for these 4 clusters in the original feature space.

ARVATO CUSTOMER SEGMENTATION REPORT



Clusters

Cluster Analysis

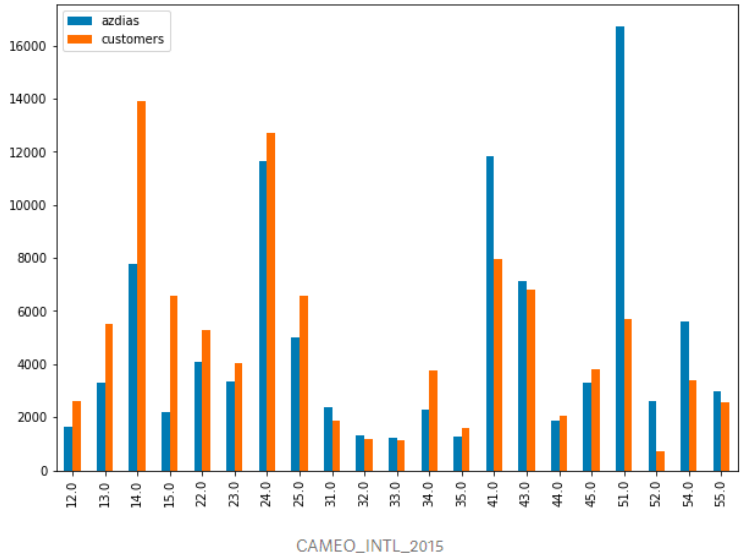
The table below shows feature values for the four cluster centers for features in which feature values greatly differ between over- and under-represented clusters. Below we discuss the differences with the aid of visuals.

FEATURES	OVERREP:3	OVERREP:8	UNDERREP:4	UNDERREP:16
AKT_DAT_KL	1	1	6	6
CAMEO_INTL_2015	4	6	16	15
D19_GESAMT_DATUM	2	5	8	8
D19_VERSAND_DATUM	2	6	9	9
HH_EINKOMMEN_SCORE	2	2	5	5
LP_FAMILIE_FEIN	9	7	1	1
LP_FAMILIE_GROB	5	4	1	1
LP_STATUS_GROB	3	3	0	0
MOBI_RASTER	3	3	0	0
VK_DHT4A	1	3	7	7
VK_DISTANZ	3	3	9	10
VK_ZG11	2	2	6	8

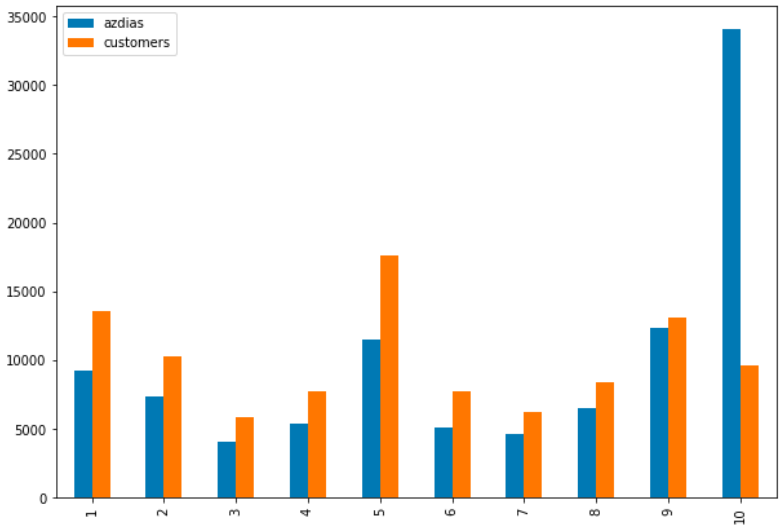
Most Important Features

- CAMEO_INTL_2015 defines household wealth. A lower value denotes greater wealth and vice versa. Customers fall into the wealthy category whereas non-customers fall into the less affluent category. The bar plot of value counts in both datasets is shown below (12 is rich and 55 is poor). Notice that the distribution for customers is more to the left. This suggests that customers, on average, belong to wealthy households as compared to the average person in Germany.

ARVATO CUSTOMER SEGMENTATION REPORT

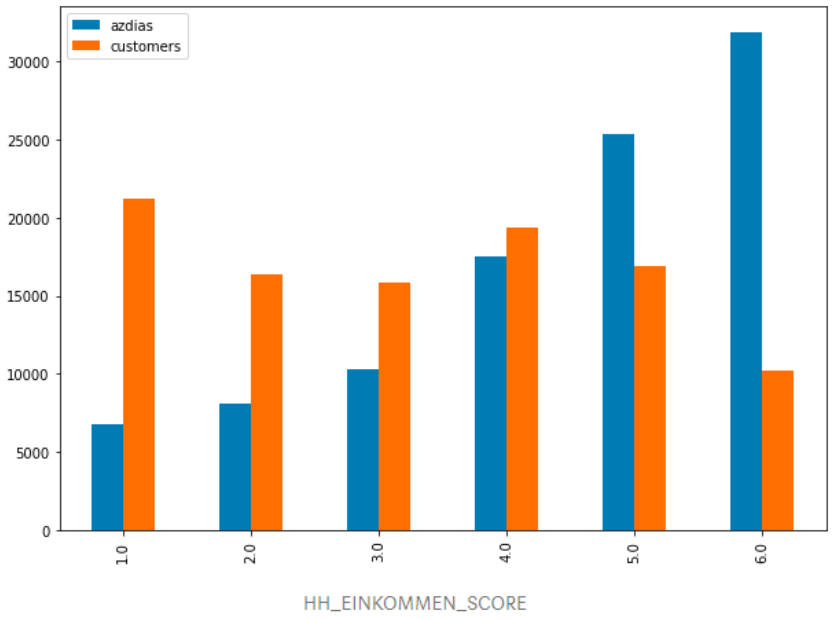


- D19_GESAMT_DATUM defines purchasing activity. A lower value denotes very high purchasing activity and vice versa. Looking at feature values in the table above, we can see that most customers have a value of 5(cluster 8) and some have a value of 2 (cluster 3). These values correspond to 'very high purchasing activity in the last 12 months' and 'slightly increased activity in the last 12 months' respectively. In comparison, non-customers have a value of 8 which corresponds to 'most recent activity of older than 2 years. Let's visualize this difference (bar plot below). The average value for customers seems to be around 5 while that for the general population is around 8. This suggests that customers, on average, have a higher purchasing power as compared to the general population**.**

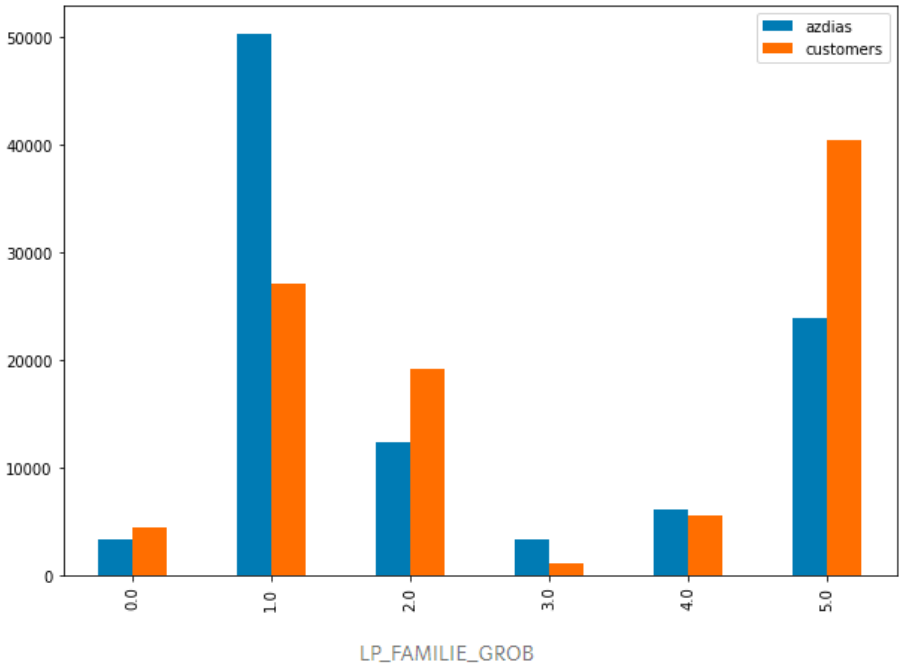


- HH_EINKOMMEN_SCORE is the estimated household net income. A lower value means greater net income and vice versa. The bar plot is shown below. Customers, on average, have a greater net income as compared to the average person in the general population. In the bar plot below, customers are centered at a value of 3 while the

general population is centered at around 5. This suggests that customers, on average, tend to have a higher net annual income.

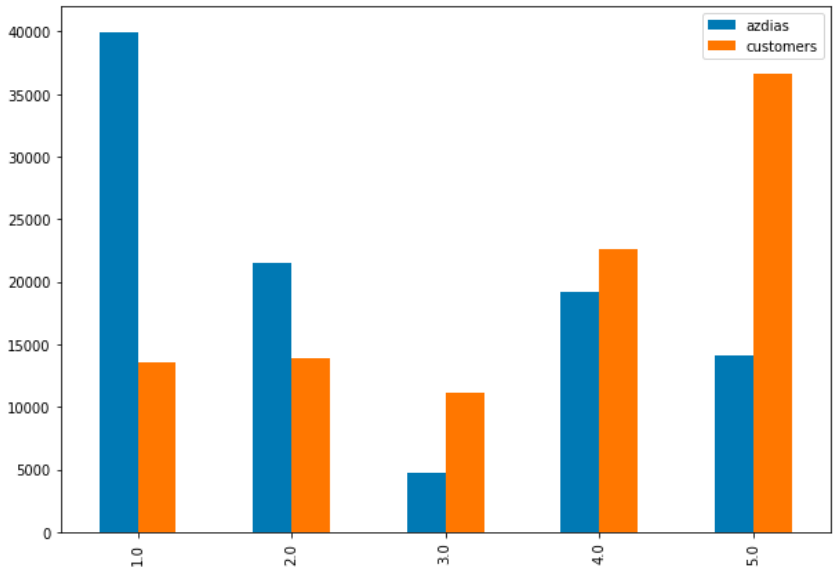


- LP_FAMILIE_GROB is family type. Lower values denote individuals living alone or with a spouse; higher values define full and multigenerational families. The bar plot is shown below. The average value for customers is more towards the right. This suggests that customers, on average, belong to full families versus people in the general population who, on average live alone or with a spouse.

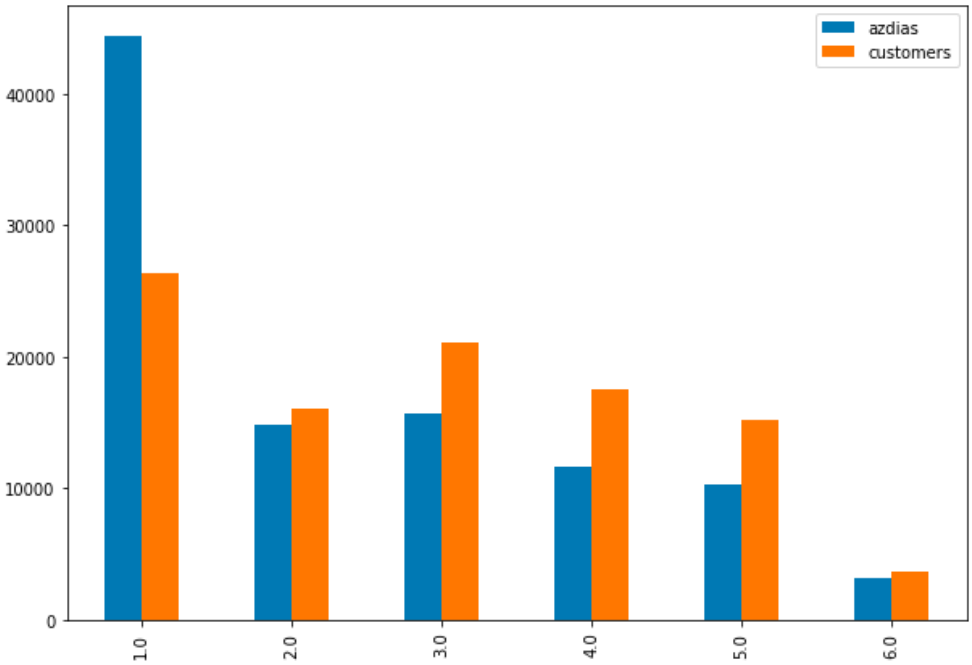


ARVATO CUSTOMER SEGMENTATION REPORT

- LP_STATUS_GROB is social status. Lower values correspond to low-income earners. Higher values correspond to houseowners and top earners. From the table above, we can see that customers have higher incomes. This suggests that customers, on average, have a higher social status than people in the general population**.**



- MOBI_RASTER defines overall mobility. Lower values correspond to high mobility and vice versa. Clusters 3 and 8 (mostly customers) have middle mobility. Clusters 4 and 16 have a very high mobility (mostly non-customers). This suggests that customers, on average, have either a stable living place or a shorter commute time to jobs**.**



ARVATO CUSTOMER SEGMENTATION REPORT

Predictive Modeling

The goal of this section is to build a predictive model that is able to predict which individuals in the general population are potential customers for the mail-order company.

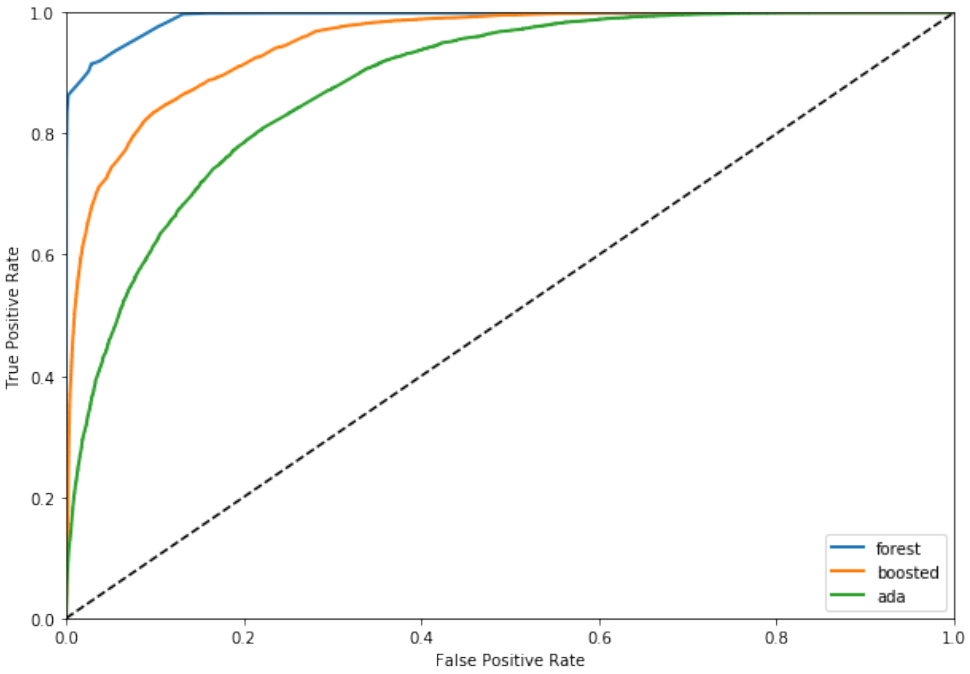
Transformations

MAILOUT_TRAIN and MAILOUT_TEST are cleaned using the already established data cleaning function. The two datasets are then transformed using the transformation pipeline that has already been constructed and which is shown in the preprocessing section. Care is taken not to reduce the number of rows for these datasets during cleaning. The parameter of 'train' is set to True in the function to avoid the issue.

There is severe class imbalance in MAILOUT_TRAIN as the negative class label (0) far exceeds the positive class label (1) in number. In order to solve this issue, oversampling is performed on the minority class. The SMOTE algorithm from imbalance-learn is used for oversampling.

Classifier Evaluation

Three ensemble methods are evaluated using cross validation on the MAILOUT_TRAIN: RandomForestClassifier, GradientBoostedCLassifier and AdaBoostedClassifier. Cross Validation is applied to evaluate each type of classifier. At each step of cross validation, training data is split into three parts. The model is trained on 2/3rds and evaluated on the other third. ROC curves for the three classifiers are shown below. The dotted line represents the ROC curve of a purely random classifier; a good classifier stays as far away from that line as possible (toward the top-left corner). RandomForestClassifier curve stays the furthest from this line. It has the highest ROC_AUC (area under the curve) score of 0.9915 as compare to GradientBoostedCLassifier (0.9494) and AdaBoostedClassifier (0.8814).



ROC curve

ARVATO CUSTOMER SEGMENTATION REPORT

Refinement

In this section, grid search is used to parameterize the RandomForestClassifier so that the final model has optimal hyperparameters. The parameter grid consists of:

{n_estimators: [100, 125, 150],
max_depth: [3, 4, 5, 6],
min_samples_split: [5,6]}

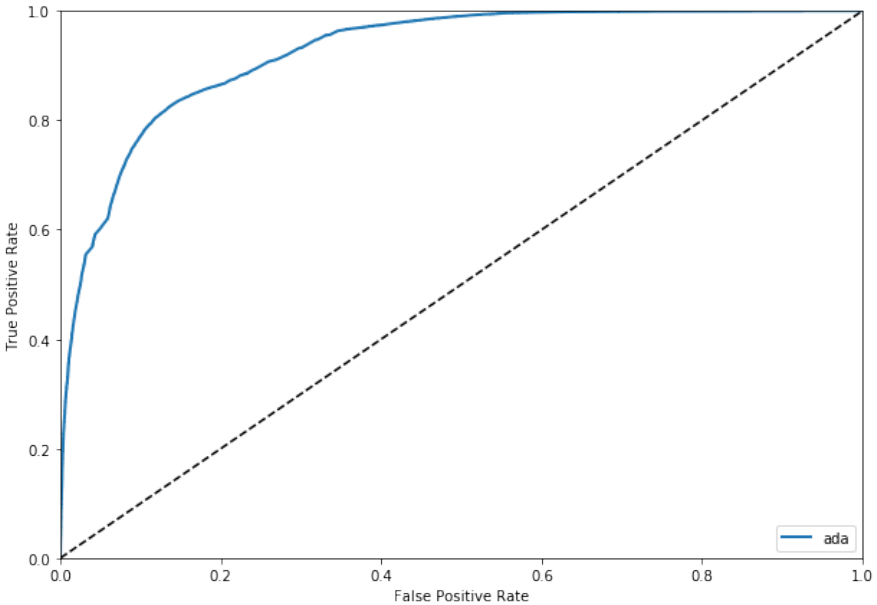
The grid search scoring is set to 'roc_auc' and number of splits in cross validation is set to 3. Hyperparameters for the best estimator are shown in the figure below. Mean cross-validated score of this best estimator is 0.9253.

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                       max_depth=6, max_features='auto', max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=5,
                       min_weight_fraction_leaf=0.0, n_estimators=125,
                       n_jobs=-1, oob_score=False, random_state=None, verbose=1,
                       warm_start=False)
```

Best Estimator

Model Evaluation

Cross validation is used to check model performance. The ROC curve is shown below. The curve is remains relatively far away from the dotted line, meaning it has an above average ROC_AUC score and has good performance.



ROC curve

ARVATO CUSTOMER SEGMENTATION REPORT

The parameterized model is applied on MAILOUT_TEST. The resulting positive class probabilities are posted to Kaggle. **The final test 'roc_auc' score is 0.73573**.

Justification

The final model is only as good as every step in the data science process. Since a lot of time was spent on data cleaning and transformation, the model has good data to be trained on. Additionally, the model is built by first selecting a viable classifier from among a bunch of classifiers and then parameterizing that model to get optimal performance. The final model outperforms other ensemble methods in terms of the ROC_AUC score. The final ROC_AUC score on Kaggle is 0.73473 so the model works relatively well.

Reflection

In this project, demographics data for customers of a mail-order company and the general population was analyzed to identify parts of the population that describe the company's core customer base. Furthermore, a predictive model was built to predict whether an individual is a potential customer for the company.

The first part of the project dealt with data preprocessing and cleaning. There are 365 demographic features in the datasets. Most of the data was ordinal and nominal. A few features were numeric. It was required to manually inspect each feature to reveal its data type. In addition, a lot of features had placeholders for nan values which had to be identified. There was missing data, lot of features were corelated and numeric features were mostly skewed. A single data cleaning function was constructed to clean features in all datasets. A transformation pipeline was created to prepare data for unsupervised learning. Three transformers were combined in a union, with each applying transformations to different data types.

In customer segmentation, an unsupervised learning pipeline was built to apply PCA and K-Means to the transformed data. It was found that the data can be decomposed into 200 principal components while retaining 95% of the variability. CUSTOMERS and AZDIAS datasets were decomposed and clustered into 17 clusters using K-Means. 2 clusters were identified that were the target customers of the company. Customers in these clusters differentiate from the general population in that they, on average, have a much higher net income, belong to wealthy families (and live within large families) and high social status.

In supervised learning, the same transformation steps were applied to the train and test sets. In addition, the train set was found to have class imbalance. SMOTE (oversample) was applied to the train set to balance out the classes. A RandomForestClassifier was chosen from among three ensemble methods due to its higher 'roc_auc' score in cross validation. The classifier was parameterized using grid search. The mean cross-validation score on the training set was 0.9253. The resulting performance of the model on Kaggle was 0.73573.

ARVATO CUSTOMER SEGMENTATION REPORT

Improvements

The manual stage of data cleaning could have been handled better. More time could have been allocated to manually checking data correlation between ordinal as well as nominal features. This could be done by constructing bar plots of features and comparing them together to see if distributions match up. Additionally, it would have been better to translate all feature names into English to get a better understanding of the features, their correlations and possibly significance.

Clustering could have been done better had a more established rule was used for determining the number of clusters instead of the elbow method. There are clustering metrics better than the elbow method. For example, the silhouette coefficient clearly shows a peak in its plot that indicates that the optimal number of clusters has been reached.

The predictive model could be developed better using some kind of bias/variance indicator plot. For example, learning curves could be used to evaluate different classifiers which in turn could lead to a better classifier selection and higher score.