Definition

Project Overview

Finding a new customer is the main objective for the corporations in across various domains e.g. insurance, retail etc. With machine learning and high computing power we can process big data and establish relations between the data and predict who will be the new customer

In this project, I analyzed demographics data for customers of a mail-order sales company in Germany, comparing it against demographics information for the general population.

Used unsupervised learning techniques to perform customer segmentation, identified the parts of the population that best describe the core customer base of the company. Then, applied supervised learning technique on a dataset with demographics information for targets of a marketing campaign for the company and use a model to predict which individuals are most likely to convert into becoming customers for the company.

The data has been provided by Bertelsmann Arvato Analytics, and represents a real-life data science task. The data includes general population dataset, customer segment data set, mailout campaign dataset with response and test dataset that needs to make predictions.

The project was inspired by below link

https://becominghuman.ai/predicting-buying-behavior-using-machine-learning-a-case-study-on-sales-prospecting-part-i-3bf455486e5d

My passion towards behavioral economics also motivated me to select this project.

Problem Statement

The goal is to find who will be the customer by comparing the existing customer base with general population; the tasks involved are the following:

1. Download general population (AZDIAS), customers (CUSTOMERS), Mailout train and test datasets

2. Data Exploration and Preprocessing
   Get to know the data and Identify missing or unknown data and impute the data accordingly.

3. Customer Segmentation using Unsupervised Learning
   Analyze general population and customer segment data sets and use supervised learning techniques to perform customer segmentation, identify the parts of the population the best describes the core customer base of the company.

   Apply Principal Component Analysis (PCA) technique for dimensionality reduction. Then, elbow curve used to identify the best number of clusters for K-means algorithm. Finally, applied K-means to make segmentation of population and customers and determine description of target cluster for the company.

4. Supervised Learning Model
   Built machine learning model using response of marketing campaign and used model to predict which individuals are most likely to convert into becoming customers for the company.

Metrics

For imbalanced classification tasks like this accuracy can give a false assumption regarding the classifier's performance, so it's better to rely on precision and recall, in the same way a precision-recall curve is better to calibrate the probability threshold in an imbalanced class scenario as a ROC Curve.

Precision-Recall curves are appropriate for imbalanced datasets so area under the curve (AUC) used as a summary of the model performance.

Source: http://www.davidsbatista.net/blog/2018/08/19/NLP_Metrics/

Analysis

Data Exploration

There are two data description excel sheets and four data files:

AZDIAS: Demographics data for the general population of Germany. It has 891211 persons data and 366 features.

	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	ALTER_KIND1	ALTER_KIND2	ALTER_KIND3	ALTER_KIND4	ALTERSKATEGORIE_FEIN	ANZ_HAUSHALTE_AKTI
0	910215	-1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	910220	-1	9.0	0.0	NaN	NaN	NaN	NaN	21.0	11.0
2	910225	-1	9.0	17.0	NaN	NaN	NaN	NaN	17.0	10.0
3	910226	2	1.0	13.0	NaN	NaN	NaN	NaN	13.0	1.0
4	910241	-1	1.0	20.0	NaN	NaN	NaN	NaN	14.0	3.0

5 rows × 366 columns

CUSTOMERS: Demographics data for customers of a mail order company. It has 191652 rows and 369 features

	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	ALTER_KIND1	ALTER_KIND2	ALTER_KIND3	ALTER_KIND4	ALTERSKATEGORIE_FEIN	ANZ_HAUSHALTE_AK
0	9626	2	1.0	10.0	NaN	NaN	NaN	NaN	10.0	
1	9628	-1	9.0	11.0	NaN	NaN	NaN	NaN	NaN	N
2	143872	-1	1.0	6.0	NaN	NaN	NaN	NaN	0.0	
3	143873	1	1.0	8.0	NaN	NaN	NaN	NaN	8.0	
4	143874	-1	1.0	20.0	NaN	NaN	NaN	NaN	14.0	

5 rows × 369 columns

MAILOUT_TRAIN: Demographics data for individuals who were targets of a marketing campaign. It has 42982 persons data and 367 features

	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	ALTER_KIND1	ALTER_KIND2	ALTER_KIND3	ALTER_KIND4	ALTERSKATEGORIE_FEIN	ANZ_HAUSHALTE_AKTIV
0	1763	2	1.0	8.0	NaN	NaN	NaN	NaN	8.0	15.0
1	1771	1	4.0	13.0	NaN	NaN	NaN	NaN	13.0	1.0
2	1776	1	1.0	9.0	NaN	NaN	NaN	NaN	7.0	0.0
3	1460	2	1.0	6.0	NaN	NaN	NaN	NaN	6.0	4.0
4	1783	2	1.0	9.0	NaN	NaN	NaN	NaN	9.0	53.0

5 rows × 367 columns

MAILOUT_TEST: Demographics data for individuals who were targets of a marketing campaign. It has 42833 persons data and 366 features

	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	ALTER_KIND1	ALTER_KIND2	ALTER_KIND3	ALTER_KIND4	ALTERSKATEGORIE_FEIN	ANZ_HAUSHALTE_AKTIV
0	1754	2	1.0	7.0	NaN	NaN	NaN	NaN	6.0	2.0
1	1770	-1	1.0	0.0	NaN	NaN	NaN	NaN	0.0	20.0
2	1465	2	9.0	16.0	NaN	NaN	NaN	NaN	11.0	2.0
3	1470	-1	7.0	0.0	NaN	NaN	NaN	NaN	0.0	1.0
4	1478	1	1.0	21.0	NaN	NaN	NaN	NaN	13.0	1.0

5 rows × 366 columns

Exploratory Visualization

The below Kolmogorov-Smirnov fit tests show how the given data is distributed for the few features

Fig. 1 Following Kolmogorov-Smirnov fit test shows how the best-ager typology feature is distributed

	General Population	Test Statistic	p-value
Best-ager typology	99	0.691344746069	0.0

Export to plot.ly »

Since our p-value is read as 0.0, we have strong evidence for not rejecting the null hypothesis that there's no difference between the means and conclude that a significant difference does exist.

Fig. 2 Following Kolmogorov-Smirnov fit test shows how the age classification through prename analysis feature is distributed

	General Population	Test Statistic	p-value
Age classification data	99	0.841344746069	0.0

Export to plot.ly »

Since our p-value is read greater than 0.0, we have strong evidence for rejecting the null hypothesis that there's a difference between the means and conclude that a significant difference does not exist.

Fig. 3 Following Kolmogorov-Smirnov fit test shows how the gender data feature is distributed

	General Population	Test Statistic	p-value
Gender Data	99	0.841344746069	0.0

Export to plot.ly »

Since our p-value is read greater than 0.0, we have strong evidence for rejecting the null hypothesis that there's a difference

between the means and conclude that a significant difference does not exist.

Fig. 4 Following Kolmogorov-Smirnov fit test shows how the transaction activity BANKS in the last 24 months data feature is distributed

	General Population	Test Statistic	p-value
Bank transactions activity	99	0.5	0.0

Export to plot.ly »

Since our p-value is read as 0.0, we have strong evidence for not rejecting the null hypothesis that there's no difference between the means and conclude that a significant difference does exist.

Fig. 5 Following Kolmogorov-Smirnov fit test shows how financial typology be prepared feature is distributed

	General Population	Test Statistic	p-value
Financial typology	99	0.841344746069	0.0

Export to plot.ly »

Since our p-value is read greater than 0.0, we have strong evidence for rejecting the null hypothesis that there's a difference between the means and conclude that a significant difference does not exist.

Algorithms and Techniques

For this customer segmentation problem, I employed 2 unsupervised learning algorithms and 1 supervised learning technique for prediction. This customer segmentation aims to find natural groupings in general population data that reveal some feature-level similarities.

Used principal component analysis (PCA) algorithm to reduce the dimensionality of general population data set. Then used k-means clustering to assign each general population group to a particular cluster based on where a group lies in component space.

Perform Dimensionality Reduction

Used sklearn PCA class to apply principal component analysis on the data, thus reducing the number of features within a dataset while retaining the "principal components".

K-means algorithm

Used the unsupervised clustering algorithm, k-means, to segment the customers data using their PCA attributes. This clustering algorithm identifies clusters of similar data points based on their component makeup.

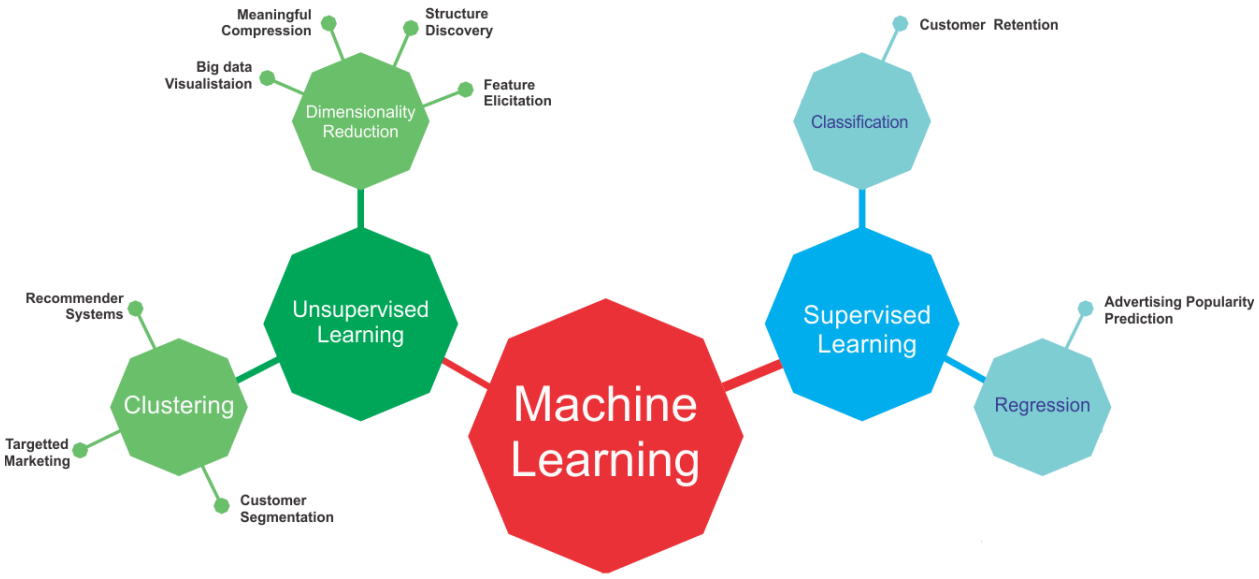
Supervised Learning for prediction

Fit a classifier using GridSearchCV with KFold, n estimators to train and test the data and find the highest ROC AUC score across multiple classifiers/regressors.

The following parameters can be tuned to optimize the model:

- Training parameters
    - Learning rate
    - Batch size
    - Training Length

Below fig depicts the complete flow.



Benchmark

To create and initial benchmark for this imbalanced classifier, I used a simple logistic regression classifier AUC score with other classifiers scores.

Methodology

Data Preprocessing

The preprocessing step consist of the following steps:

1. Load general population AZDIAS and customers data
2. Load and Prepare missing or unknown values based on given DIAS Attributes – Values 2017.xlsx.
3. Identify rows and columns which has missing or unknown data
4. Asses columns types

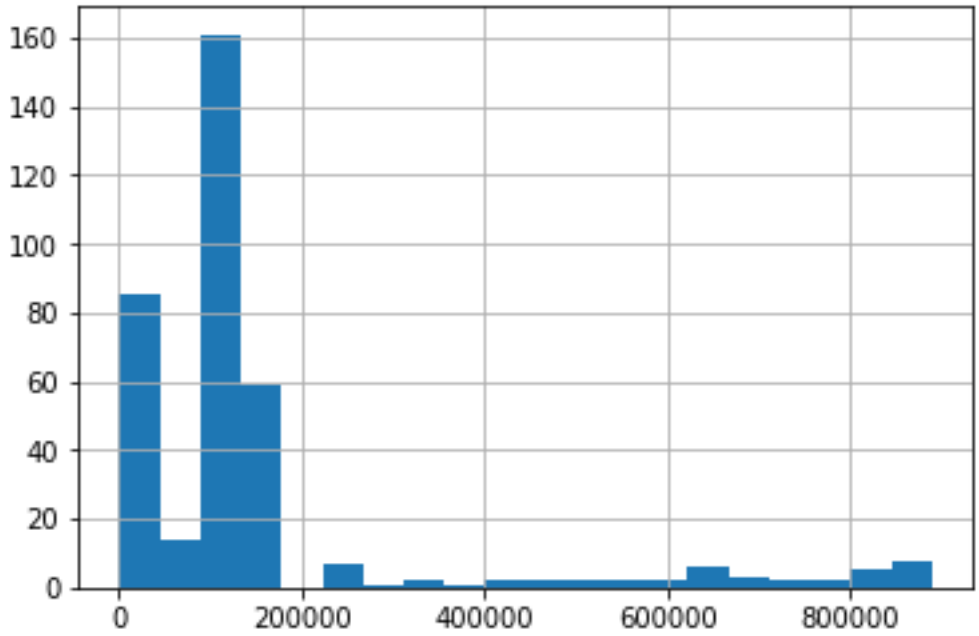
5. Prepare a clean function which removes and adds dummy columns
6. Impute the values for missing data
7. Apply feature scaling
8. Apply feature engineering e.g. Indicator variables (wealthy customer, age limit, type of cars, creating dummy variables etc.)

I created a data frame of missing or unknown values based on DIAS Attributes – Values 2017.xlsx.

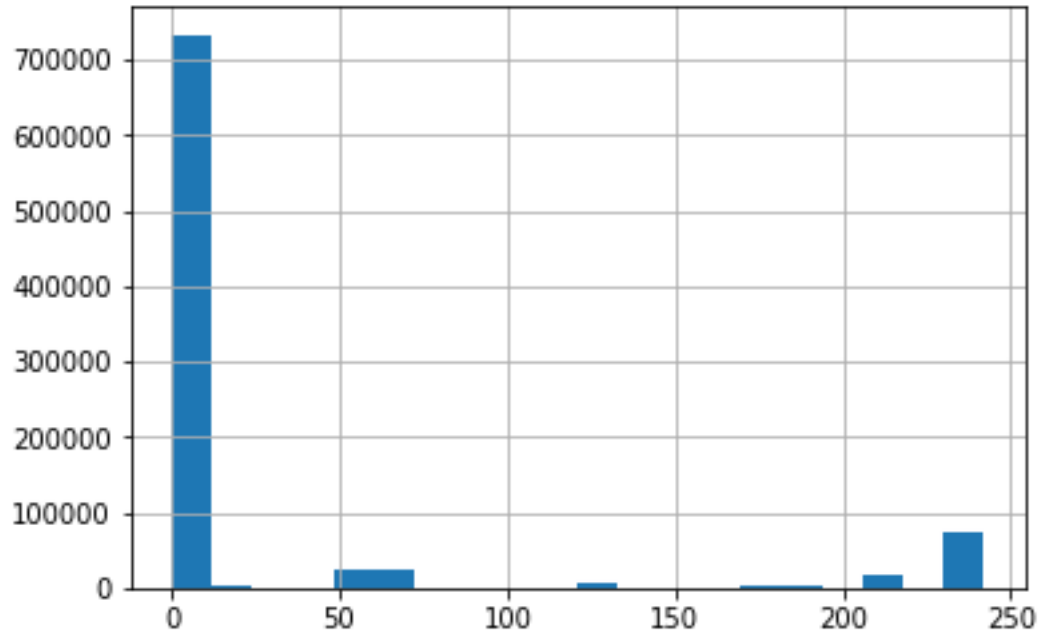
Prepared a missing or unknown (-1 or 0) columns list from the DIAS attributes sheet then compared with AZDIAS dataset features. Next find out the values that correspond to missing value codes of AZDIAS dataset and converted all of the missing or unknowns to NaNs.

After analyzing the data, found that there are lot of rows and columns/features has missing data. The analysis identified that the columns with NaNs which were more than 200k and rows with NaNs which were more than 10 per row. Since these doesn’t add any value for the analysis, these attributes were dropped from analysis.

Number of columns with NaNs



Number of Rows with NaNs



Additionally, there are other columns that were dropped based on the following reasons:

1. Column with unique values e.g LNR
2. Columns with more than 10 categories
3. Missing data e.g MIN_GEBAEUDEJAHR

By going through the DIAS attributes list, identified columns which are categorical e.g gender type

Crated dummy variables for columns with less than 10 unique values for simplicity. Adjusted feature OST_WEST_KZ data appropriately

At the end 732489 rows and 360 columns left.

After feature engineering, imputed missing values using mean method so that data doesn't have any more NANs.

Finally applied feature scaling so that principal component vectors are not influenced by the natural differences in scale for features.

Implementation

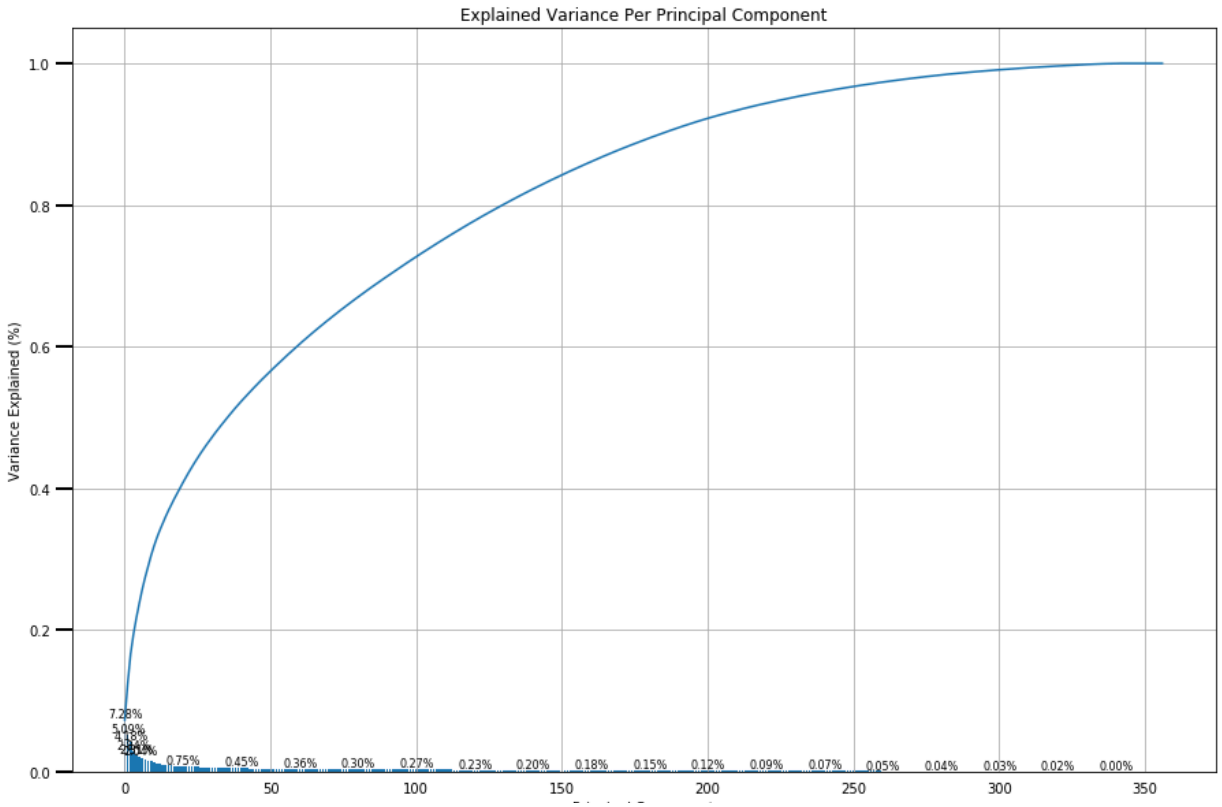
The implementation process can be split into two main stages:

1. The customer segmentation
2. The Prediction

During the 1st stage, I have good data and ready to be fed into machine learning model. Each data point has 360 features, which means the data is 360 dimensional and an algorithm like K means has difficulties and result is often noisier clusters.

To address this issue, I applied dimensionality reduction technique called Principal Component Analysis (PCA) on the dataset which will find the optimal number of components which captures the greatest amount of variance of data.

I plotted explained variance bar plot (see below). There is a visual reduction in explained variance after ~200 components. This number of transformed features results in 90% explained variance. So, 200 transformed features were retained for the clustering part of the project.



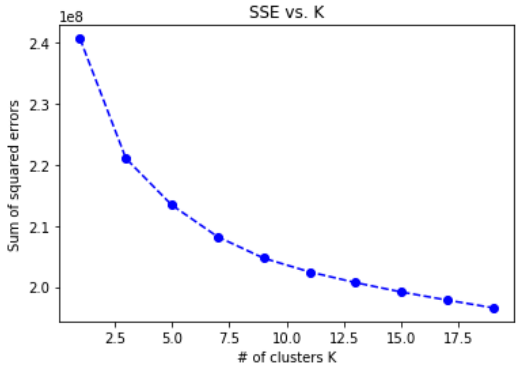
K-Means Clustering

In this step, I used k-means clustering to view the top PCA components. To do this, first fit these top principal components to the k-means algorithm and determine the best number of clusters. Determining the ideal number of clusters for k-means model can be done by elbow graph.

Elbow graph method was used to identify an ideal number of clusters for K-means clustering on the PCA-transformed data.

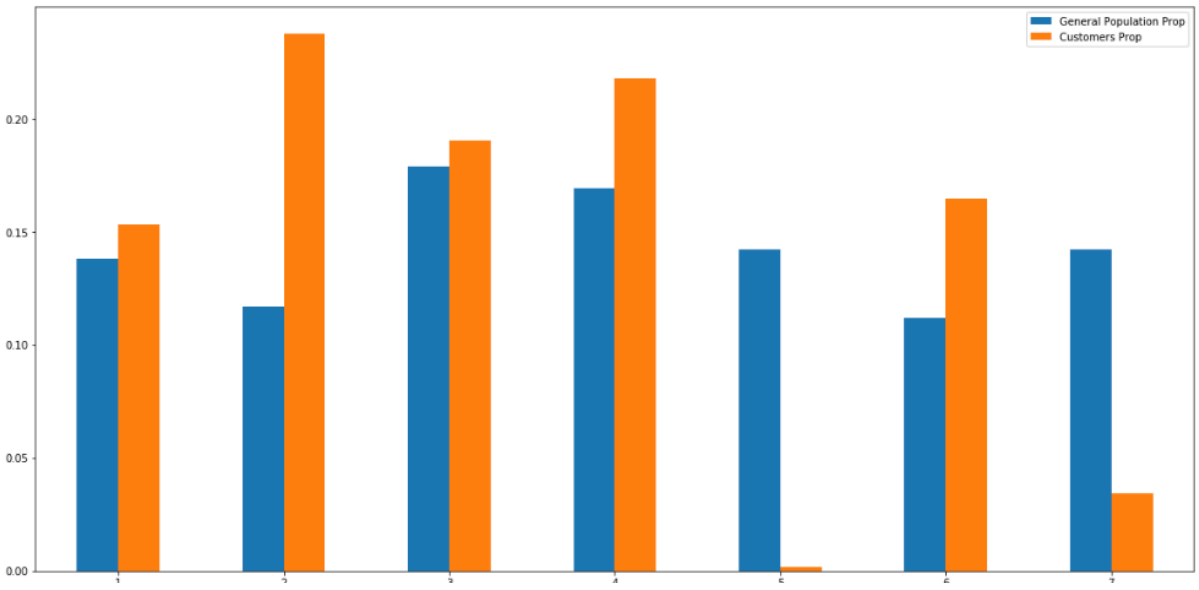
Average of sum of squared errors (SSE) within-cluster distances was plotted against number of clusters.

The Elbow graph below shows that after 7 clusters at (the elbow) the change in the value of inertia is no longer significant and most likely, neither is the variance of the rest of the data after the elbow point. We can discard everything after k=7 and proceed with next step.



Visualize and Interpret the Clusters

Then unsupervised learning technique K-means clustering fit and predict methods applied to the general population and customers data. Then compare the proportion of data in each cluster for the customer data to the proportion of data in each cluster for the general population as shown in the below figure.



Clusters with the highest negative difference in proportion between customers and general audience are underrepresented in the customer data (no interest clusters #5 and #7).

The ability to notice otherwise unseen patterns and to come up with a model to generalize these patterns on to observations is precisely why the tools like PCA and k-means are essential.

Supervised Learning Model

During 2nd stage, it is clear that which parts of the population are more likely to be customers of the mail-order company and build a prediction model. Loaded train data and implemented pre-processing steps similar to customers dataset. Used various classifiers to train the dataset and select the model which has the highest ROC AUC score.

Refinement

As mentioned in the Benchmark section, model performance will be measured based on ROC AUC scores across multiple classifiers.

To get the initial result, a simple logistic regression classifier was used; and the result of ROC AUC score is around 0.65

Fit a classifier using GridSearchCV and calculated ROC AUC scores with various classifiers and captured the scores.
Out of many iterations, GradientBoosting Classifier achieved highest ROC AUC classifier score of 0.7

The model performance was further improved upon by using the following technique:

- Algorithm changed to SAMME.R
- Increased learning rate
- Increased estimators

Results

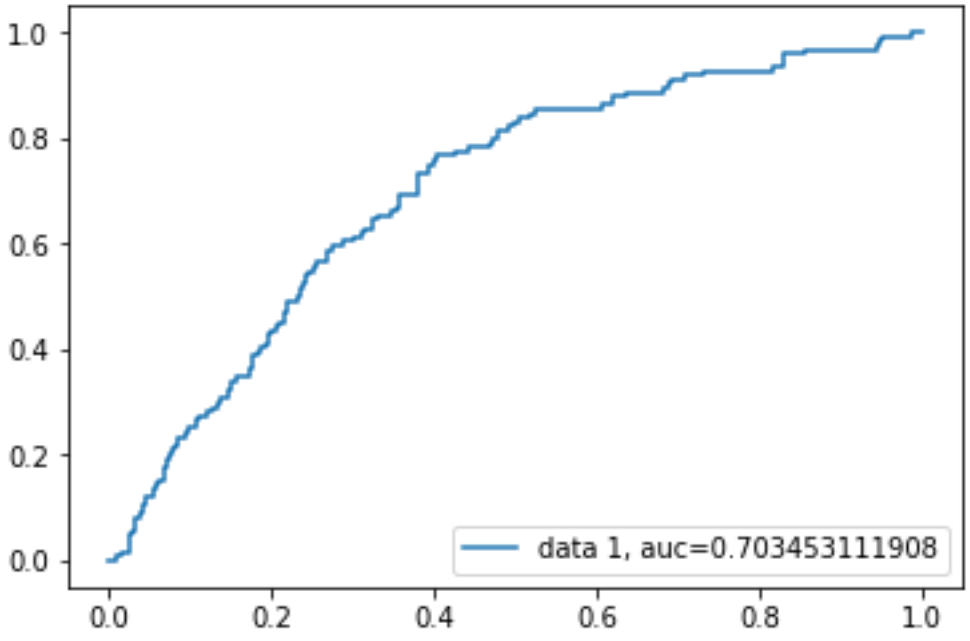
Model Evaluation and Validation

When working with classification problems, AUC (Area Under the Curve) ROC (Receiver Operating Characteristics) curves are one of the most important evaluation metrics for checking the model's performance.

This curve is a measurement of the performance for a classification problem at various thresholds settings. ROC is a probability curve and AUC represent degree or measure of separability and informs of how much the model is capable of distinguishing between classes.

The final architecture and hyperparameters were chosen because they performed the best among the tried combinations with the KFold Cross Validation models and random states.

GradientBoostingRegressor outperformed all other classifiers and regressors. It is observed that if a small change in the dataset or learning rate may not affect the model performance.




```
0.703453111908
```

Fig: ROC AUC score

As it can be observed in the previous graph, our curve provides, for the chosen model, values of TPR (True Positive Rate) higher than FPR (False Positive Rate) at any stage of the curve and an area of 0.7, which means there is a good separability degree, our model is performing pretty well and no overfitting is being produced.

Justification

The results of customer segmentation and ROC AUC scores proved that this model can be used for real world scenarios for any industry domain.

Conclusion

Reflection

The process used for this project can be summarized using the following steps:

- In the first part the assessment and preprocessing of the data was performed. There were 366 columns to analyze and not all of them had description. There were identified a lot of missing values and missing information about attributes. Feature engineering, Feature selection and handled missing data process was created that was further utilized in supervised and unsupervised parts.
- In the unsupervised part, the dimensionality reduction using PCA was performed to 200 latent features that describe 90% of explained variance. K-means clustering to 7 clusters identified 2 clusters that are target customers of the company. These are share of midclass cars in the PLZ8 and 20–25 years of age.
- Lastly, Gradient Boosting Classifier was used and parameterize to build supervised model and make predictions over testing dataset on KAGGLE. The resulted performance of supervised learning algorithm is 70%.

This project potentially has some improvements. For example, there are other ways to preprocess the data: choose another threshold for dropping rows and columns, MICE, choose different transformations for the columns, apply MinMax Scaler instead of Standard Scaler, impute data in another way.

Improvement

Improvement of supervised model can be tested by using PCA dimensionality reduction. We could also choose attributes that have the most difference in clustering for overrepresented and underrepresented data and build supervised model using only these attributes. And also increase the hyper parameters for better optimization.