CAPSTONE REPORT

Machine Learning Engineer Nanodegree

DATA SCIENCE

Customer Segmentation Report for Arvato Financial Services

Shalaby, Mohamed

Udacity

Contents

Domain Background

I have chosen Customer Segmentation Report for Arvato Financial Services, Arvato is an internationally active services company that develops and implements innovative solutions for business customers from around the world. These include SCM solutions, financial services and IT services, which are continuously developed with a focus on innovations in automation and data/analytics. Globally renowned companies from a wide variety of industries – from telecommunications providers and energy providers to banks and insurance companies, e-commerce, IT and Internet providers – rely on Arvato's portfolio of solutions. Arvato is wholly owned by Bertelsmann. The services business also includes the global customer experience company Majorel, which is listed on Euronext Amsterdam and in which Bertelsmann holds a stake of around 40 percent.

Source to the Domain Background from arvato Company site

Timeline of process management and the improvement steps:

1- **Customer Segmentation Report Use unsupervised learning methods to analyze attributes of established customers and the general population in order to create customer segments.**

2- **Supervised Learning Model Within this part, a third dataset is provided with attributes from targets of a mail order campaign, and the previous analysis will be used to build a machine learning model that predicts whether or not each individual will respond to the campaign.**

3- **Kaggle Competition Once the model is chosen, predictions on the campaign data are calculated as part of a Kaggle Competition.**

Problem Statement

In this project, we will analyze demographics data for customers of a mail-order sales company in Germany, comparing it against demographics information for the general population. We'll use unsupervised learning techniques to perform customer segmentation, identifying the parts of the population that best describe the core customer base of the company. Then, we'll apply what we've learned on a third 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.

Datasets and Inputs

- \`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).

- Two metadata files:

1- DIAS Information Levels - Attributes 2017.xlsx: top level list of attributes and descriptions.

2- DIAS Attributes - Values 2017.xlsx: details for each feature.

- **In addition to helper file "function.py" to clean our datasets and evaluate ML algorithms.**

Solution Statement

The goal of Arvato project is to acquire new customers from the German population by predicting who would becoming a customer from demographic profiles of customers. Each row of the demographics files represents a single person, but also includes information outside of individuals, including information about their household, building, and neighborhood.

- **The platform we'll use to approach our solution is The Jupyter Notebook which is a web-based interactive computing platform. The notebook combines live code, equations, narrative text, visualizations. We'll use libraries like numpy, pandas, matplotlib, scikit-learn and Autogluon AutoML library. Python version 3.9.13 | packaged by conda-forge.**

library	version
numpy	**1.23.2**
pandas	**1.4.3**
Matplotlib	**3.5.3**
Sklearn	**1.1.2**
autogluon	**0.5.1**

Steps of our solution approach:

1- **Customer Segmentation Report:**
We'll begin the project by using unsupervised learning methods k-means clustering algorithm to analyze attributes of established customers and the general population in order to create customer segments. We'll use the information from the first two files to figure out how customers ("CUSTOMERS") are similar to or differ from the general population at large ("AZDIAS").

2- **Supervised Learning Model:**
Then we'll use our previous analysis to build a machine learning model to make predictions on the other two files ("MAILOUT"), predicting which recipients are most likely to become a customer for the mail-order company. The machine learning models we'll used are Logistic Regression, AdaBoost Classifier, Random Forest Classifier, SGD Classifier, Gradient Boosting Classifier and Autogluon AutoML library.

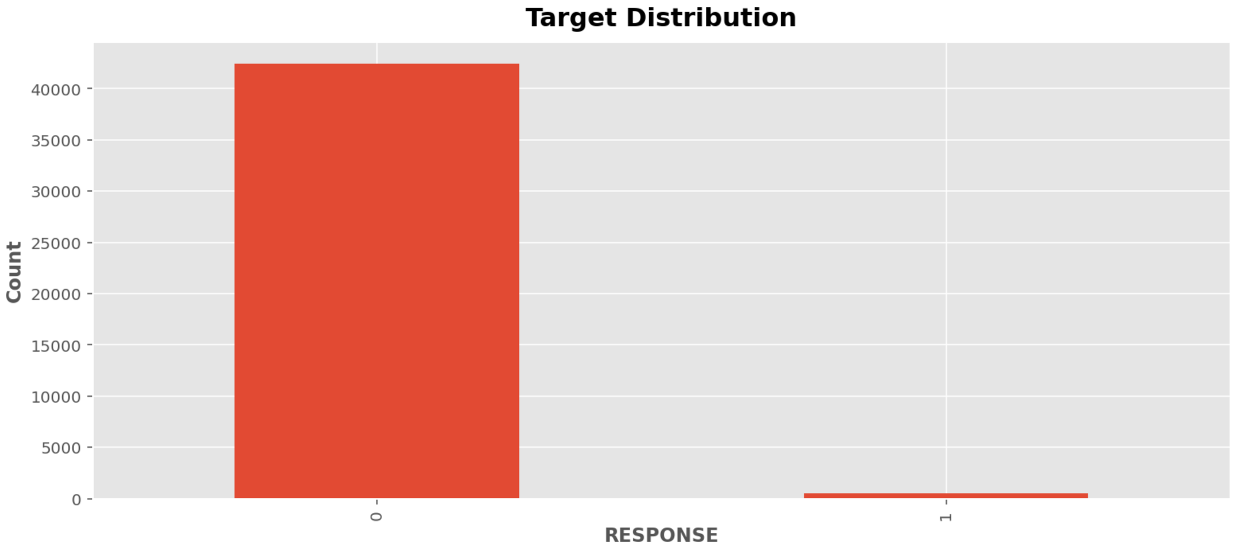
Benchmark Model

A benchmark model for our binary classification problem would be a Logistic Regression Model. The performance of this model can be used to compare against different algorithms to help choose one supervised algorithm over another.

Evaluation Metrics

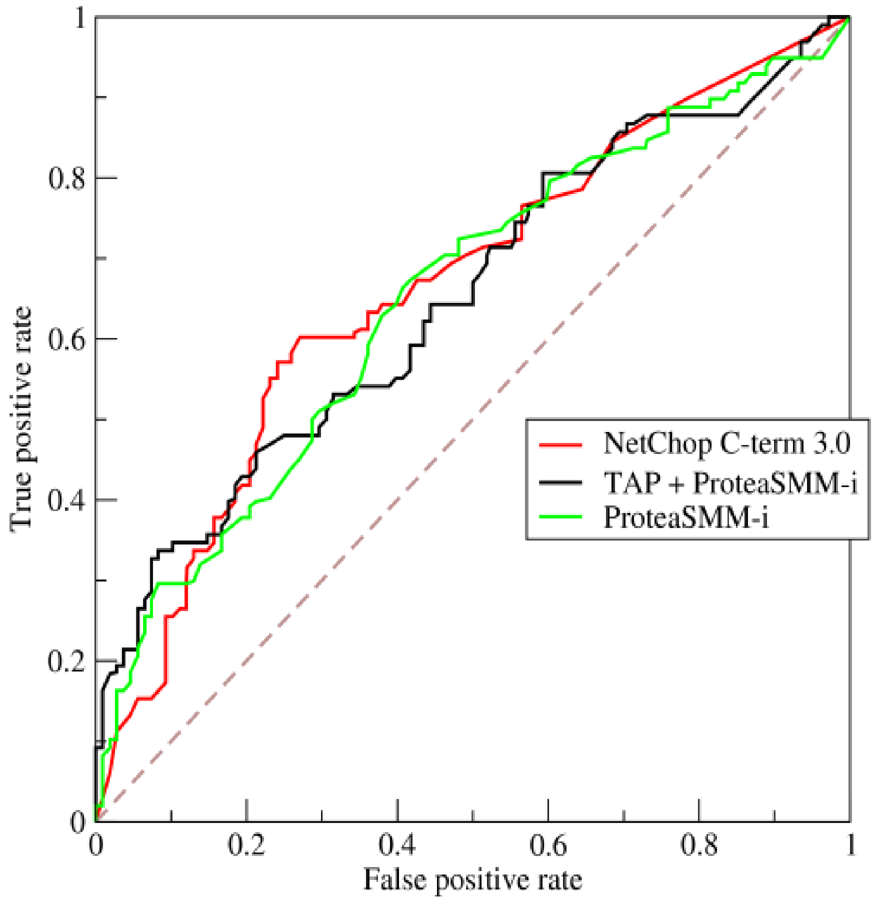
Our target is "RESPONSE" it determines if customer "1" or not "0"

Due to the great imbalance in our target as shown below, we will use AUC for the ROC curve as metric.



And also, the evaluation metric for the Kaggle competition is AUC for the ROC curve.

Picture to our metric is shown below.



The selected model will be used to make predictions on the mailout campaign data in competition through Kaggle. We will select the model that make best score according to our evaluation metric. References for our metric.

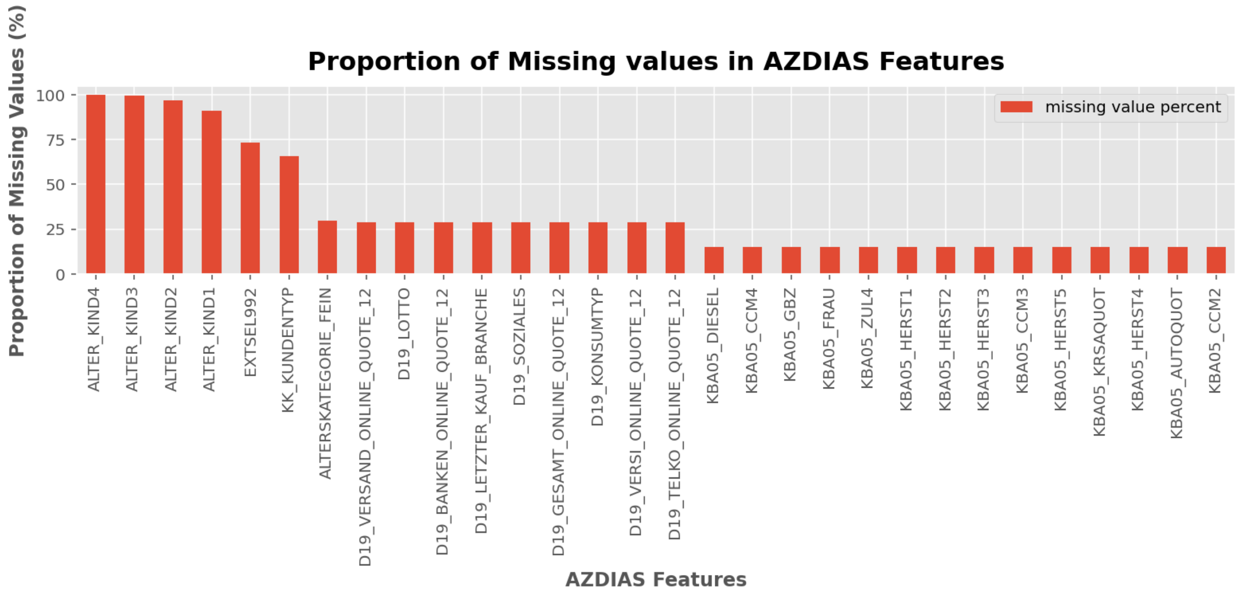
Project Design

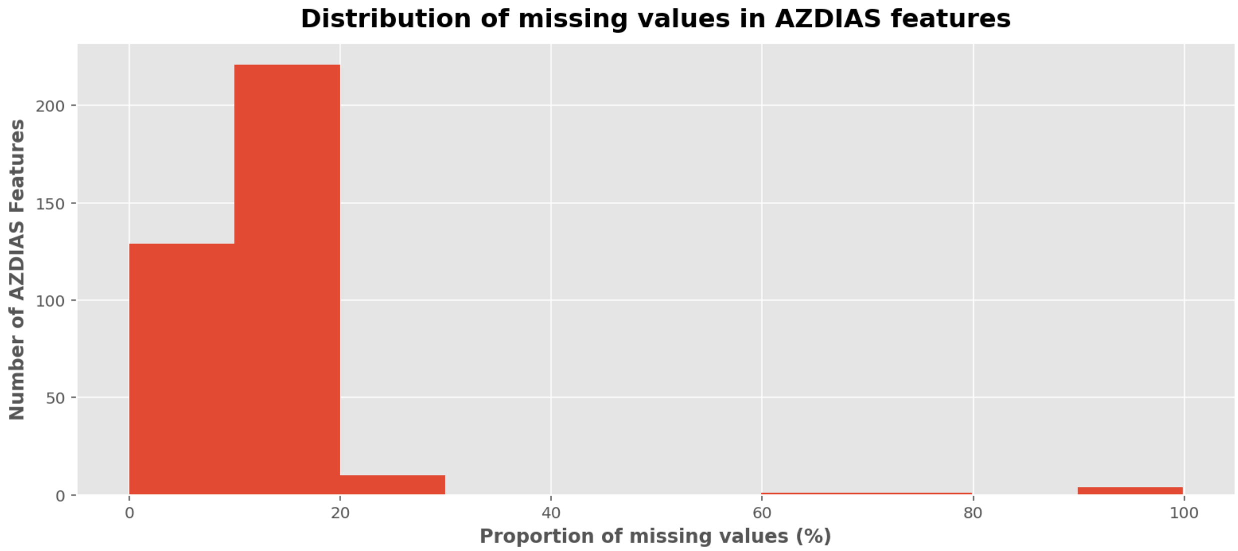
1- **Data assessment and cleaning**

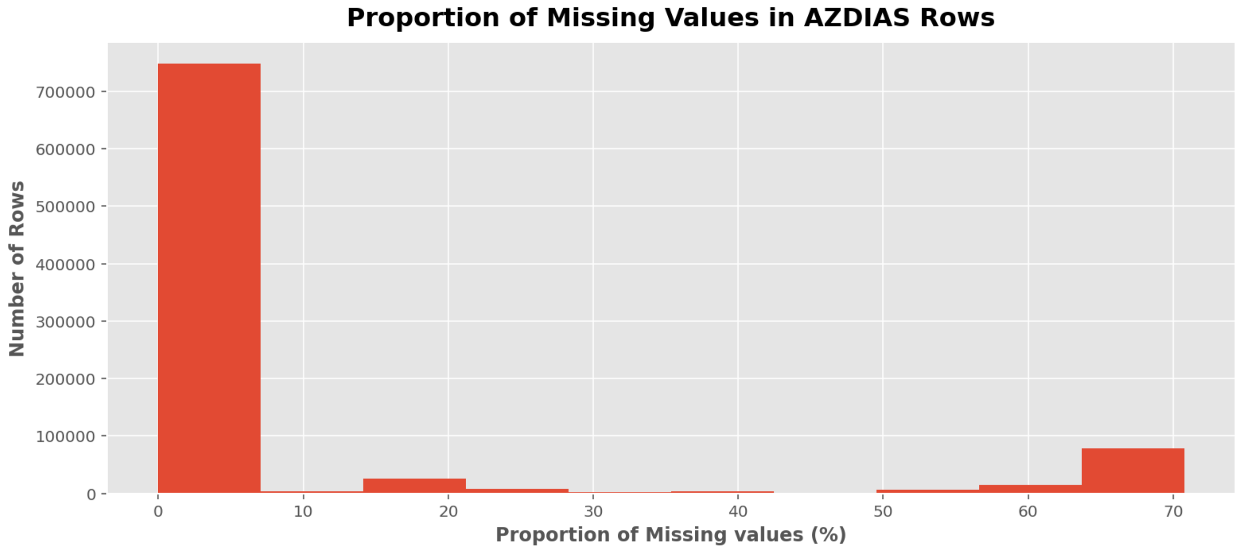
- **To know and clean our data and be familiar with data types and null values in our dataset. We will use our helper function.py to clean our datasets.**
  - ✓ **We will Convert 'OST_WEST_KZ' column to 1 for 'W' and 0 for 'O'.**
  - ✓ **Dropping most twenty null columns.**
  - ✓ **Dropping Correlated Features more than 0.9.**
  - ✓ **Dropping null rows more than 10% except for "mailout_test" dataset.**
  - ✓ **Define global variable drop_cols in function.py file to use it for the following datasets after cleaning azdias data.**

2- **Data Visualization**

- ✓ **To identify distribution patterns in the data.**
- ✓ **To identify null values in our columns and rows.**
  - ▪ **Examples are shown below from our project.**







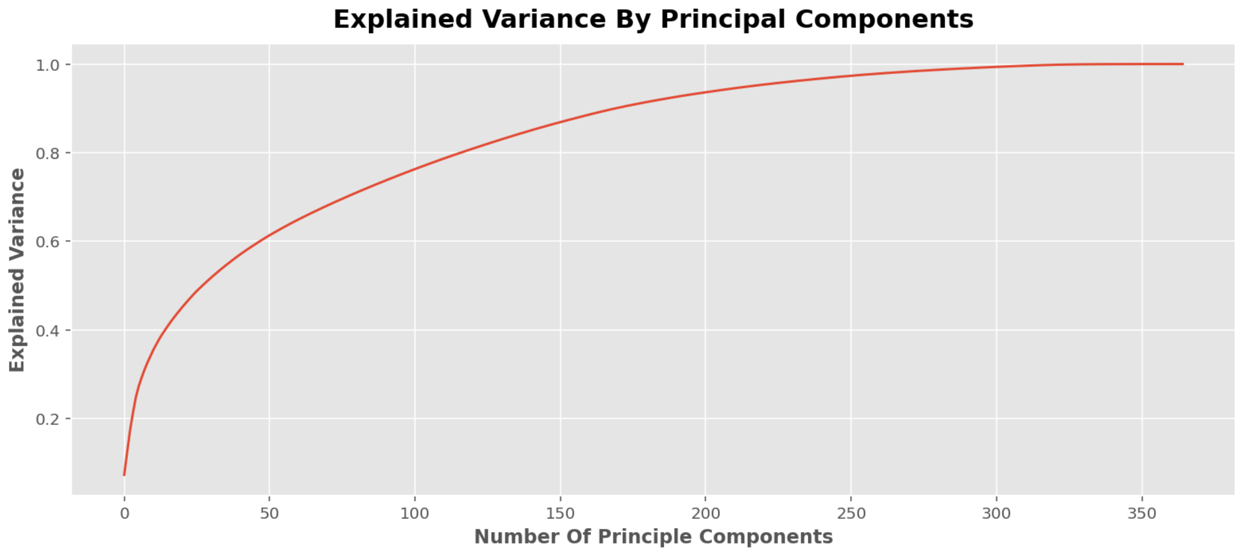
3- Feature Engineering

- **Impute missing values.**
- **One Hot encoding – Before training on categorical features, we need to one hot encode them so that we can feed numerical values to ML model.**
- **Scaling and fix data skewness in continuous variable. If there is too much skewness, then normalize through log transformation.**

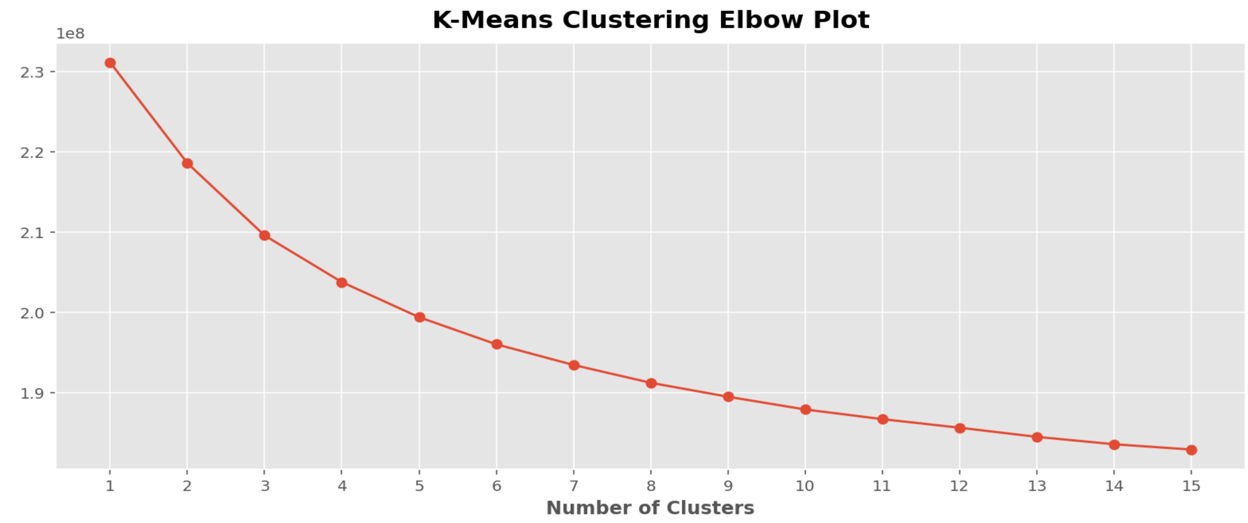
- **Dimensionality Reduction to generalize and simplify our model.**

4- Model Selection and Training

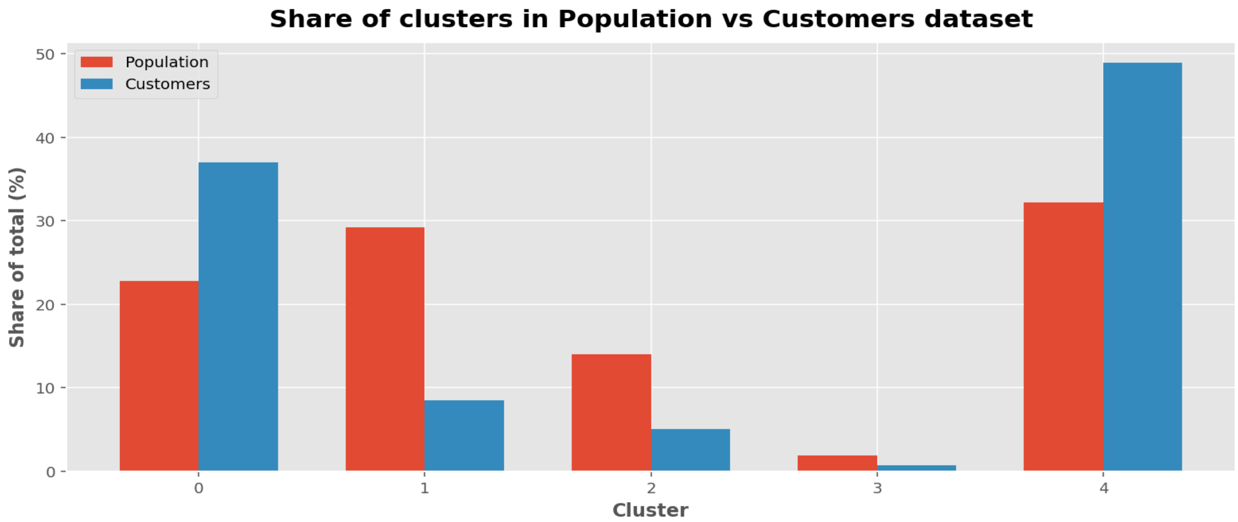
- **Customer Segmentation using unsupervised learning techniques k-means clustering algorithm.**
    - ✓ **Fit PCA object to modified azdias dataset.**
    - ✓ **plot explained variance and choose 217 components explain 95% of the variance.**



- ✓ **cluster the population dataset by applying k-means cluster algorithm.**
- ✓ **plot k-means and determine five clusters by elbow technique.**

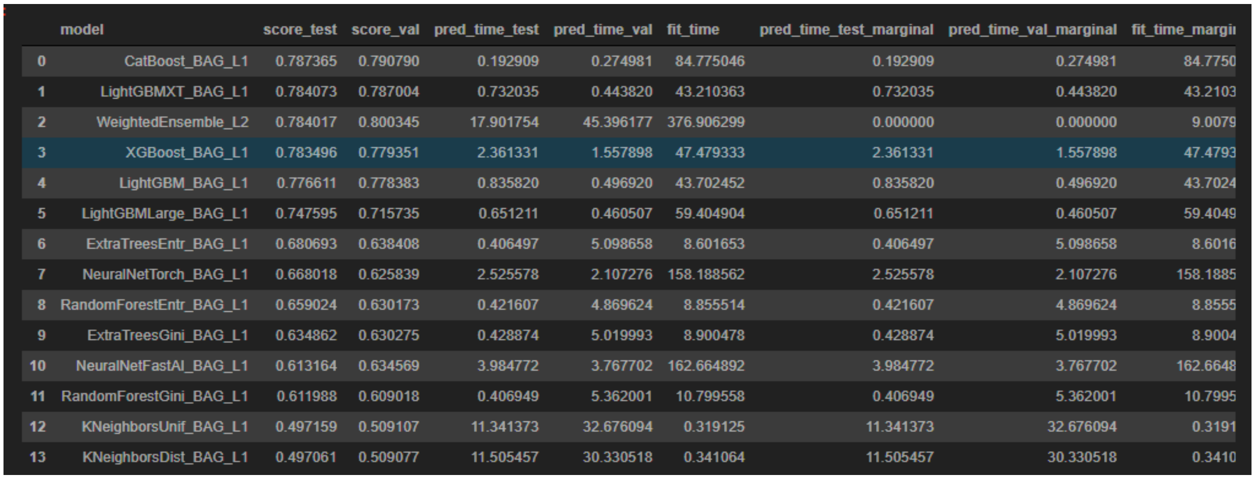


- ✓ **After Plotting cluster shares we conclude that People in Clusters 0 and 4 are more likely to be our future customers and marketing campaign should focus on these groups. While clusters 1, 2 and 3 seems to be less interesting.**



- **Supervised learning algorithms will be trained and evaluated on predicting new customers.**
  - **Logistic Regression**
  - **AdaBoost Classifier**

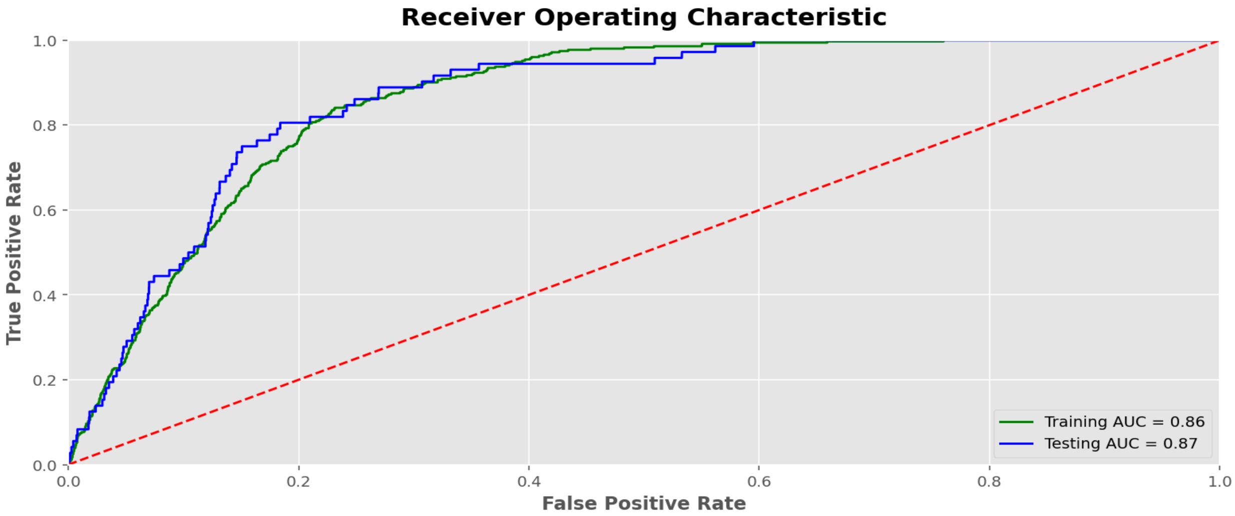
- **Random Forest Classifier**
- **SGD Classifier**
- **Gradient Boosting Classifier**
- **Autogluon AutoML library**



	model	score_test	score_val	pred_time_test	pred_time_val	fit_time	pred_time_test_marginal	pred_time_val_marginal	fit_time_margi
0	CatBoost_BAG_L1	0.787365	0.790790	0.192909	0.274981	84.775046	0.192909	0.274981	84.7750
1	LightGBMXT_BAG_L1	0.784073	0.787004	0.732035	0.443820	43.210363	0.732035	0.443820	43.2103
2	WeightedEnsemble_L2	0.784017	0.800345	17.901754	45.396177	376.906299	0.000000	0.000000	9.0079
3	XGBoost_BAG_L1	0.783496	0.779351	2.361331	1.557898	47.479333	2.361331	1.557898	47.4793
4	LightGBM_BAG_L1	0.776611	0.778383	0.835820	0.496920	43.702452	0.835820	0.496920	43.7024
5	LightGBMLarge_BAG_L1	0.747595	0.715735	0.651211	0.460507	59.404904	0.651211	0.460507	59.4049
6	ExtraTreesEntr_BAG_L1	0.680693	0.638408	0.406497	5.098658	8.601653	0.406497	5.098658	8.6016
7	NeuralNetTorch_BAG_L1	0.668018	0.625839	2.525578	2.107276	158.188562	2.525578	2.107276	158.1885
8	RandomForestEntr_BAG_L1	0.659024	0.630173	0.421607	4.869624	8.855514	0.421607	4.869624	8.8555
9	ExtraTreesGini_BAG_L1	0.634862	0.630275	0.428874	5.019993	8.900478	0.428874	5.019993	8.9004
10	NeuralNetFastAI_BAG_L1	0.613164	0.634569	3.984772	3.767702	162.664892	3.984772	3.767702	162.6648
11	RandomForestGini_BAG_L1	0.611988	0.609018	0.406949	5.362001	10.799558	0.406949	5.362001	10.7995
12	KNeighborsUnif_BAG_L1	0.497159	0.509107	11.341373	32.676094	0.319125	11.341373	32.676094	0.3191
13	KNeighborsDist_BAG_L1	0.497061	0.509077	11.505457	30.330518	0.341064	11.505457	30.330518	0.3410

5- <u>Model Testing and Predictions</u>

The best model will be used to make predictions on the test data. AdaBoost Classifier was pretty good as shown below.

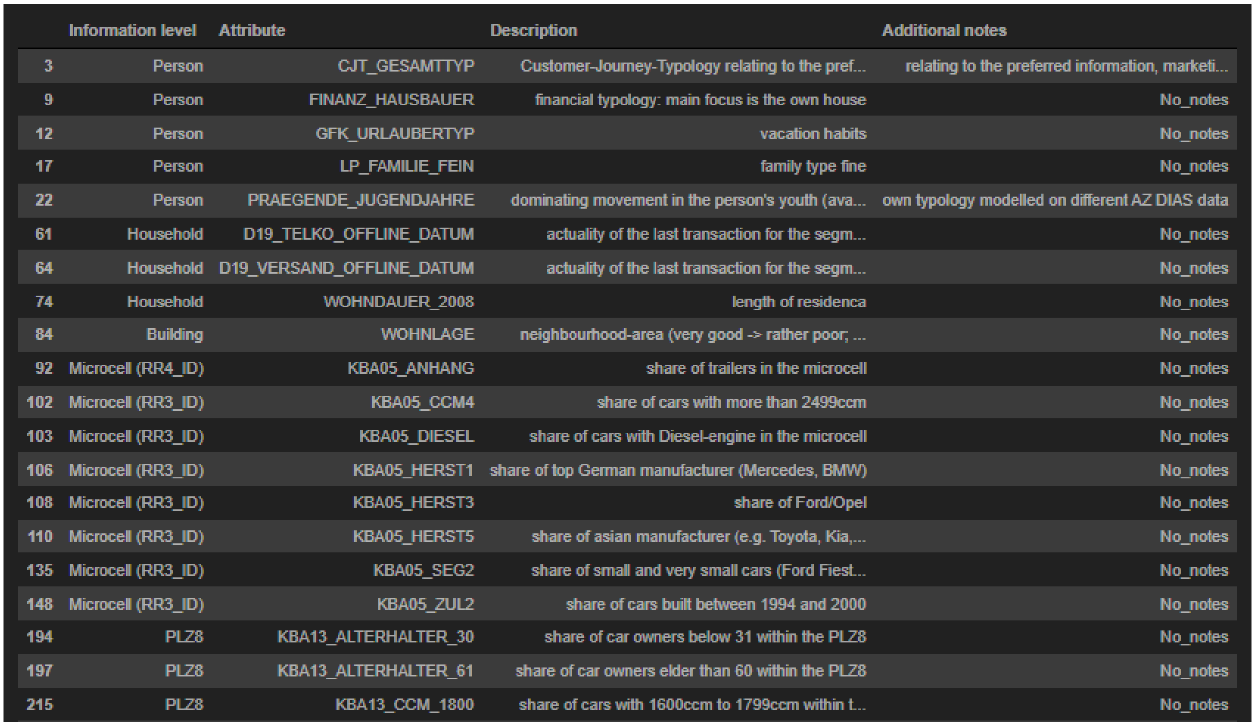


Prediction on test dataset were submitted to Kaggle. **Kaggle Submission**

Conclusion

After finishing our project, I concluded from the analysis process that We can focus on the following features and try to improve our data about population according to them to enhance our model and prediction.

Features are 46 and some of them shown below.



	Information level	Attribute	Description	Additional notes
3	Person	CJT_GESAMTTYP	Customer-Journey-Typology relating to the pref...	relating to the preferred information, marketi...
9	Person	FINANZ_HAUSBAUER	financial typology: main focus is the own house	No_notes
12	Person	GFK_URLAUBERTYP	vacation habits	No_notes
17	Person	LP_FAMILIE_FEIN	family type fine	No_notes
22	Person	PRAEGENDE_JUGENDJAHRE	dominating movement in the person's youth (ava...	own typology modelled on different AZ DIAS data
61	Household	D19_TELKO_OFFLINE_DATUM	actuality of the last transaction for the segm...	No_notes
64	Household	D19_VERSAND_OFFLINE_DATUM	actuality of the last transaction for the segm...	No_notes
74	Household	WOHNDAUER_2008	length of residenca	No_notes
84	Building	WOHNLAGE	neighbourhood-area (very good -> rather poor; ...	No_notes
92	Microcell (RR4_ID)	KBA05_ANHANG	share of trailers in the microcell	No_notes
102	Microcell (RR3_ID)	KBA05_CCM4	share of cars with more than 2499ccm	No_notes
103	Microcell (RR3_ID)	KBA05_DIESEL	share of cars with Diesel-engine in the microcell	No_notes
106	Microcell (RR3_ID)	KBA05_HERST1	share of top German manufacturer (Mercedes, BMW)	No_notes
108	Microcell (RR3_ID)	KBA05_HERST3	share of Ford/Opel	No_notes
110	Microcell (RR3_ID)	KBA05_HERST5	share of asian manufacturer (e.g. Toyota, Kia,...	No_notes
135	Microcell (RR3_ID)	KBA05_SEG2	share of small and very small cars (Ford Fiest...	No_notes
148	Microcell (RR3_ID)	KBA05_ZUL2	share of cars built between 1994 and 2000	No_notes
194	PLZ8	KBA13_ALTERHALTER_30	share of car owners below 31 within the PLZ8	No_notes
197	PLZ8	KBA13_ALTERHALTER_61	share of car owners elder than 60 within the PLZ8	No_notes
215	PLZ8	KBA13_CCM_1800	share of cars with 1600ccm to 1799ccm within t...	No_notes

And I always believe that the most important part in our Machine Learning process is data analysis and features engineering.