Capstone Project Proposal

Machine Learning Engineer Nanodegree

Vinicius Gomes Pereira

25/03/2020

The Market Segmentation Problem

Market segmentation consists of identifying a group of individuals who have characteristics in common, in a heterogeneous market. This should be seen as a powerful tool for auxiliary departments of marketing, design and strategic areas.

Several characteristics of consumers and potential consumers can be taken into account for this segmentation task, for example:

1. Demographic, geographic, social and economic criteria.
2. Personality and lifestyle criteria.
3. Product behavior criteria

The question is what is the advantage of segmentation? The market is heterogeneous and therefore it is advisable to segment it. Thus, attitudes towards these segments can be different. Segmentation allows to rationalize the means to reach a given product segment, adjusting it to the prices and costs of distribution and communication, with a view to achieving balance. It also allows a specialization of the company playing with the strategic variables - price, product, distribution and communication - avoiding waste.

Problem Statement

Consider a company's marketing campaign (Arvato Financial Services), in which we need to select those individuals who can become the company's future customers. For this task, we have the following databases: demographic information from Germany (country where the company is located) and information from individuals who are already customers of this company.

We used unsupervised machine learning algorithms to analyze the market and segment it, selecting the main characteristics that can best describe the company's consumer.

Then, we used this information to create a predictive model that can determine with reasonable precision whether an individual can be a likely consumer or not, when subjected to a particular marketing campaign.

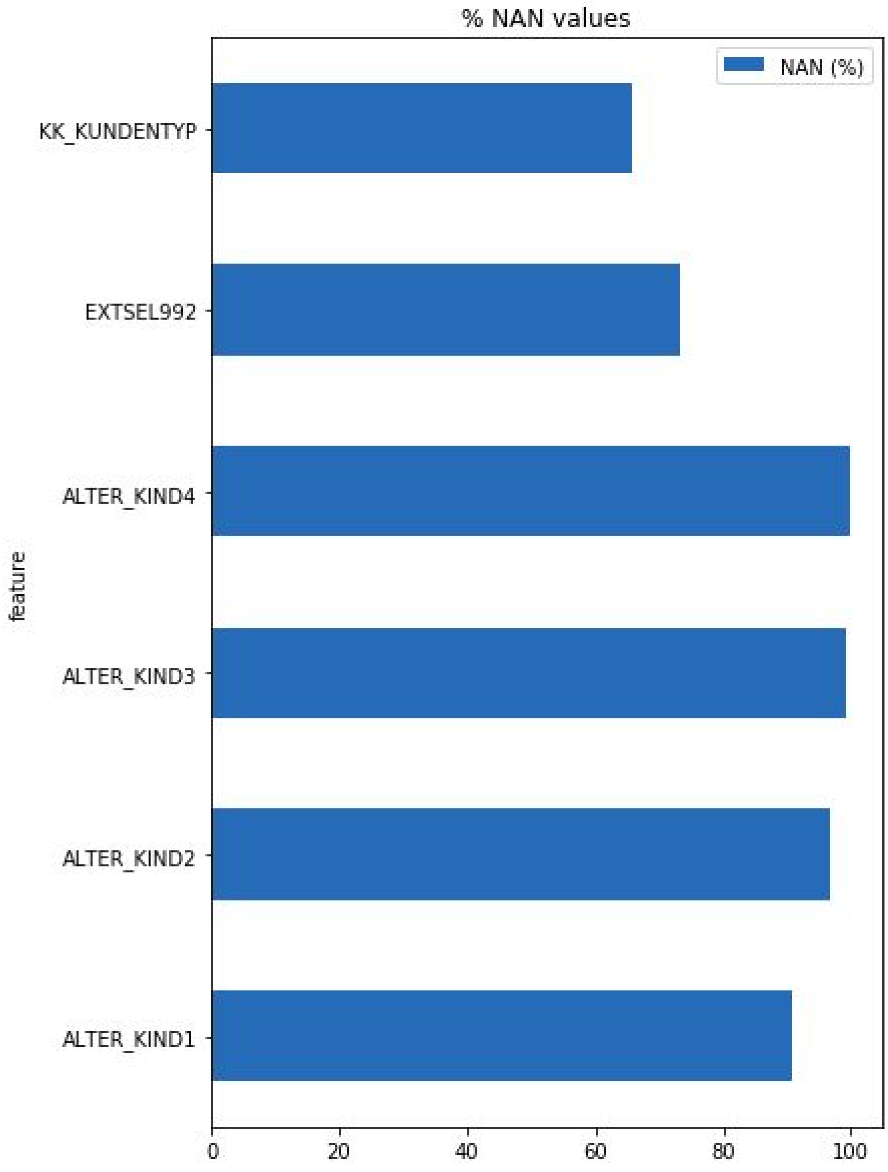
Datasets and Inputs

There are four data files associated with this project, that were provided by Udacity :

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

The first and the second file will be used in the customer segmentation step. The third one will be used to train the predictive model that will be used to identify possible consumers. The forth file will be used to measure how well our model performs in the classification task.

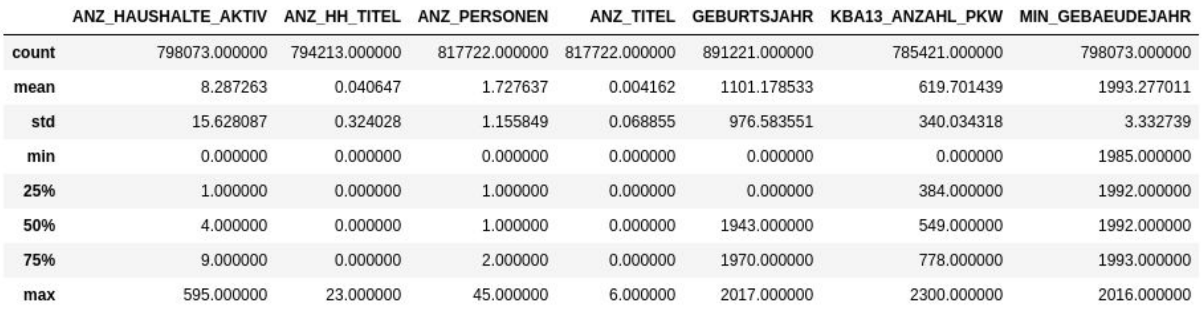
The data set Udacity_AZDIAS_052018 has 891211 rows and 366 features. Some features has high number of empty values. In the graph below, it is possible to see some features with more than 50% of the data with empty values.



Most of the features of Udacity_AZDIAS_052018 is categorical, only ANZ_HAUSHALTE_AKTIV, ANZ_HH_TITEL, ANZ_PERSONEN, ANZ_TITEL, GEBURTSJAHR, KBA13_ANZAHL_PKW, MIN_GEBAEUDEJAHR are of numeric types. The meaning of each numerical feature is described in the next table:

Numeric Features	
Feature	Description
ANZ_HAUSHALTE_AKTIV	number of households in the building
ANZ_HH_TITEL	number of academic title holder in building
ANZ_PERSONEN	number of adult persons in the household
ANZ_TITEL	number of professional title holder in household
GEBURTSJAHR	year of birth
KBA13_ANZAHL_PKW	number of cars in the PLZ8
MIN_GEBAEUDEJAHR	year the building was first mentioned in our database

The next table has some information of theses features, like mean, standard deviation, min and max values and percentiles. It is possible to notice , for example that a 1.72 is the average number of adults in the households and 1985 is the oldest year of a building mentioned in database.



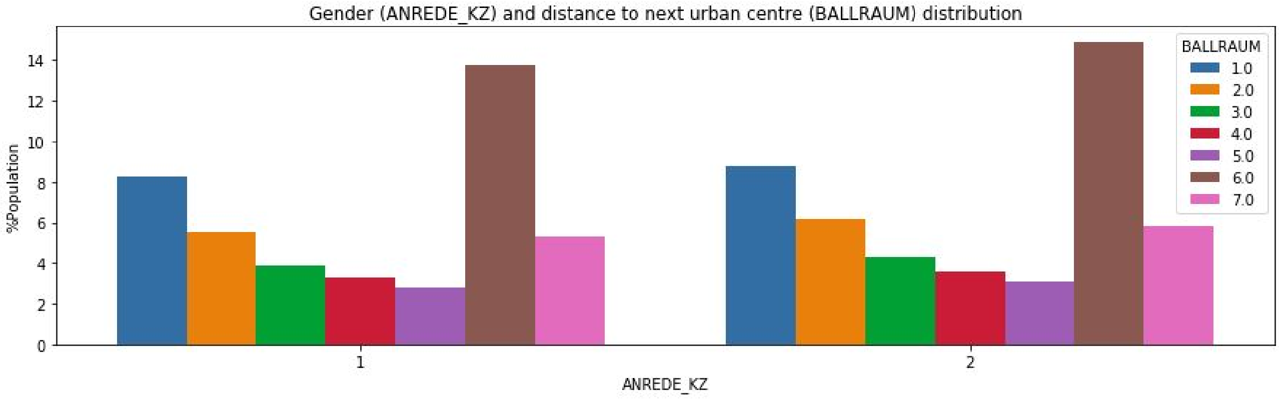
	ANZ_HAUSHALTE_AKTIV	ANZ_HH_TITEL	ANZ_PERSONEN	ANZ_TITEL	GEBURTSJAHR	KBA13_ANZAHL_PKW	MIN_GEBAEUDEJAHR
count	798073.000000	794213.000000	817722.000000	817722.000000	891221.000000	785421.000000	798073.000000
mean	8.287263	0.040647	1.727637	0.004162	1101.178533	619.701439	1993.277011
std	15.628087	0.324028	1.155849	0.068855	976.583551	340.034318	3.332739
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1985.000000
25%	1.000000	0.000000	1.000000	0.000000	0.000000	384.000000	1992.000000
50%	4.000000	0.000000	1.000000	0.000000	1943.000000	549.000000	1992.000000
75%	9.000000	0.000000	2.000000	0.000000	1970.000000	778.000000	1993.000000
max	595.000000	23.000000	45.000000	6.000000	2017.000000	2300.000000	2016.000000

Some features are categorical, for example:

Categorical Features	
Feature	Description
ALTERSKATEGORIE_GROB	age classification through prename analysis
ANREDE_KZ	gender
BALLRAUM	distance to next urban centre

D19_BANKEN_ONLINE_QUOTE_12	amount of online transactions within all transactions in the segment bank
D19_BUCH_RZ	transactional activity based on the product group BOOKS and CDS
D19_KONSUMTYP	consumption type
D19_KK_KUNDENTYP	consumption movement in the last 12 months

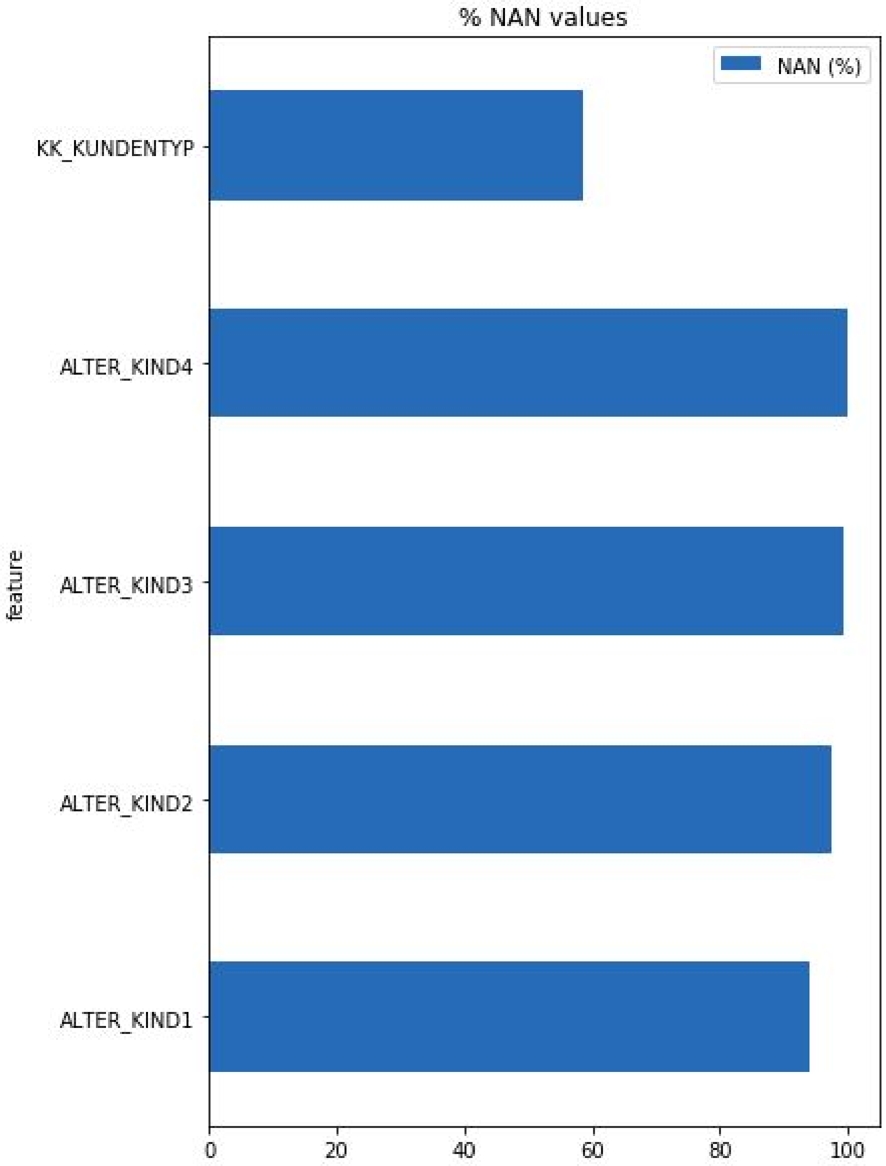
In the graph below, we can see how people are distributed from the urban centre, by gender, where de categorical value (ANREDE_AZ) 1 means ‘Male’ and 2 means ‘Female’ and BALLRAUM values are explained in the next table.



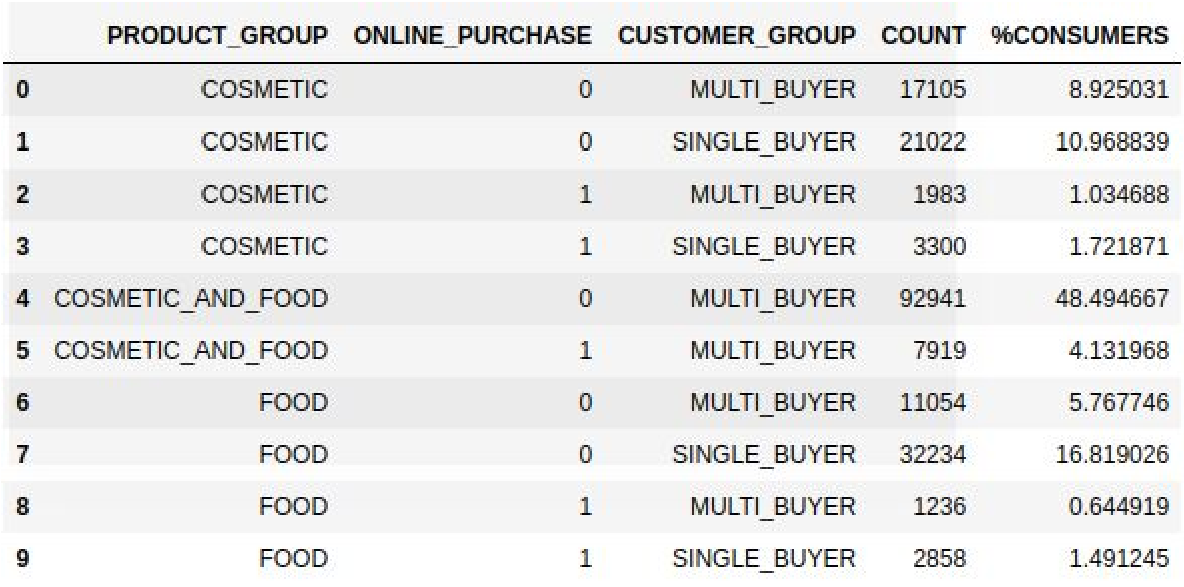
BALLRAUM	
Value	distance to next urban centre
-1	unknown
1	till 10 km
2	10 - 20 km
3	20 - 30 km
4	30 - 40 km
5	40 - 50 km
6	50-100 km
7	more than 100 km

Hence, we can conclude that most of the people lives between 50-100 km from the next urban centre and there are not significance difference of behavior analyzing this feature by gender.

The Udacity_CUSTOMERS_052018.csv Demographics data for customers of a mail-order company with 191 652 persons with 369 features. Some features has high number of empty values. In the graph below, it is possible to see some features with more than 50% of the data with empty values. The result is similar to the dataset of general population of Germany, as we previously have discussed.



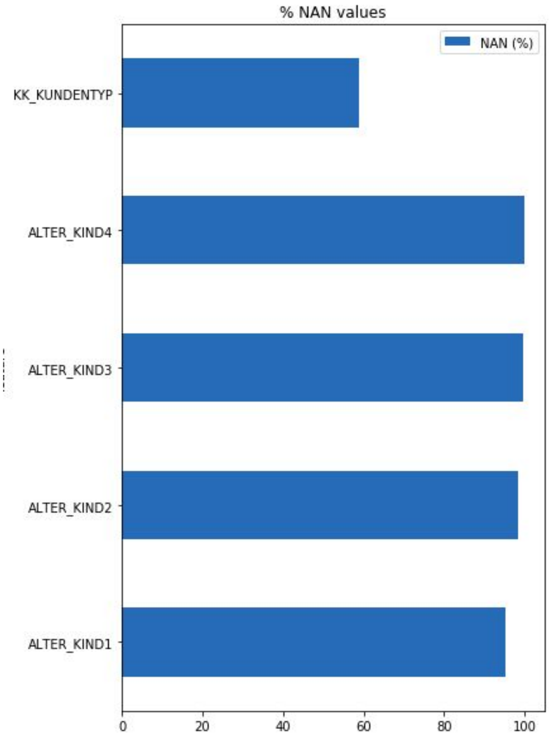
Besides, in this dataset, there are three categorical features that the first dataset does not have: ‘PRODUCT_GROUP', 'ONLINE_PURCHASE' and 'CUSTOMER_GROUP', with information about the person behavior as a consumer. The next table has the dataset of consumers grouped in theses three features:



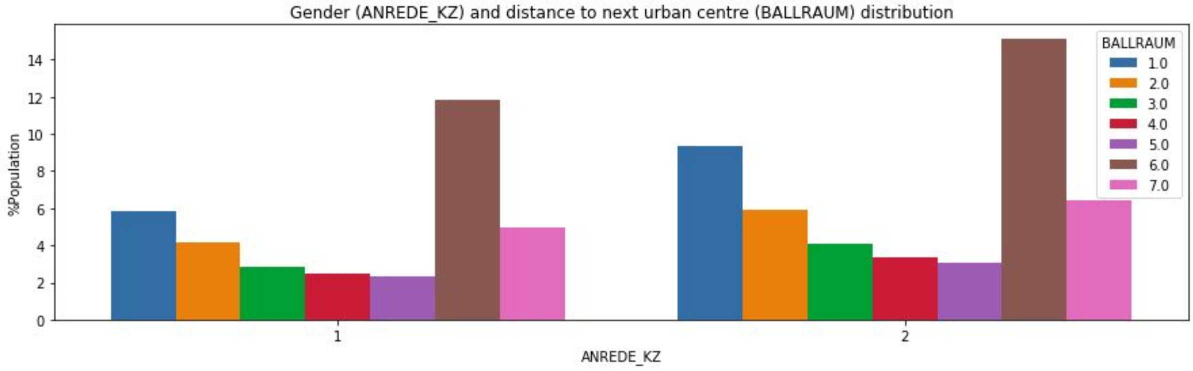
	PRODUCT_GROUP	ONLINE_PURCHASE	CUSTOMER_GROUP	COUNT	%CONSUMERS
0	COSMETIC	0	MULTI_BUYER	17105	8.925031
1	COSMETIC	0	SINGLE_BUYER	21022	10.968839
2	COSMETIC	1	MULTI_BUYER	1983	1.034688
3	COSMETIC	1	SINGLE_BUYER	3300	1.721871
4	COSMETIC_AND_FOOD	0	MULTI_BUYER	92941	48.494667
5	COSMETIC_AND_FOOD	1	MULTI_BUYER	7919	4.131968
6	FOOD	0	MULTI_BUYER	11054	5.767746
7	FOOD	0	SINGLE_BUYER	32234	16.819026
8	FOOD	1	MULTI_BUYER	1236	0.644919
9	FOOD	1	SINGLE_BUYER	2858	1.491245

In both data sets, we will perform the exploration data analysis, using PCA and K-Means.

Udacity_MAILOUT_052018_TRAIN.csv has demographics data for individuals who were targets of a marketing campaign with 42982 persons and 367 features. There is one additional column in relation to Udacity_AZDIAS_052018, with the variable response. Some features has high number of empty values. In the graph below, it is possible to see some features with more than 50% of the data with empty values.



In the next graph is plotted how people are distributed from the urban centre, by gender. It is quite similar to the first data set of population of Germany. We can conclude , for example, that most of the people lives between 50-100 km from the next urban centre and there are not significance difference of behavior analyzing this feature by gender.



The data set Udacity_MAILOUT_052018_TEST.csv has demographics data for individuals who were targets of a marketing campaign with 42 833 persons and 366 features in columns. This is the data set that we will perform classification.

Solution Statement

For Customer Segmentation Report, we will do the exploratory data analysis on the datasets Udacity_AZDIAS_052018, Udacity_CUSTOMERS_052018 and Udacity_MAILOUT_052018_TRAIN, applying unsupervised machine learning techniques like K-means with some dimensionality reduction method (PCA, for example), extracting important features from the data.

After that, for the Classification Task, we are going to use the extracted features from the prior analysis and apply some classification algorithms (like Random Forest, SVM, Logistic Regression, Decision Trees and XgBoost), using Udacity_MAILOUT_052018_TRAIN for training, separating some part of the dataset for validation.

Finally, the developed algorithm will be applied to the test dataset and submitted on Kaggle's platform. Then, we are going to deploy on AWS and make the model available for using through a web service.

Benchmark Model

First we are going to compare in the classification task our developed model, using the features that we extracted, to Naive Bayes Model using the original features of the dataset. Then, we are going to compare our model to the Kaggle ranking of this project [1].

Evaluation Metrics

Accuracy, recall and precision are common metrics to use for classification task. These three evaluation metrics are defined below, where TP, TN, FP, FN mean true positives, true negatives, false positive and false negatives in the classification process:

accuracy = (TP+ TN)/(TP + FP + TN+FN)
recall = (TP)/(TP+FN)
precision = (TP)/(TP+FP)

Besides, we are going to also evaluate the AUC for the ROC curve defined in [2]. The ROC curve is a graphic used to plot the true positive rate TPR (true positive rate) against FPR (false positive rate).

For PCA, we will compute the number of components that retains 99% of the variance of the original data. For the k-Means algorithm, we will compute the average of distances to each center cluster, and we will study the increase of the number of clusters and the decreasing of these average to decide the number of clusters of our analysis.

Project Design

We are going to follow the next steps:

1) Exploratory Data Analysis of demographics data for the general population of Germany;
2) Exploratory Data Analysis for the consumers dataset;
3) Feature Extraction, using Principal Component Analysis (PCA);
4) K-Means for customer segmentation, using the features extracted;
5) Develop classification models on training dataset, using hyperparameter tuning and validation techniques,
6) Choose the best model and submit on Kaggle platform;
7) Deploy the chosen model on AWS SageMaker;
8) Enable an endpoint for the the Web Service, using (API Gateway on AWS and Lambda Functions);

References

[1] Kaggle.com. (2020). *Udacity+Arvato: Identify Customer Segments | Kaggle*. [online] Available at: https://www.kaggle.com/c/udacity-arvato-identify-customers/leaderboard [Accessed 1 Mar. 2020].

[2] En.wikipedia.org. (2020). *Receiver operating characteristic*. [online] Available at: https://en.wikipedia.org/wiki/Receiver_operating_characteristic#Area_under_the_curve
[Accessed 1 Mar. 2020].