MACHINE LEARNING ENGINEER NANODEGREE

Bertelsmann/Arvato

Capstone proposal

Domain background :

Bertelsmann is a German private multinational conglomerate based in Gütersloh, North Rhine-Westphalia, Germany. It is one of the world's largest media conglomerates and also active in the service sector and education.

Arvato is an international service provider.
Arvato offers services, for example, in the areas of Customer Relationship Management (CRM), Supply Chain Management (SCM) and finance, as well as information technology.

source : Wikipedia

Customer segmentation can be practiced by all businesses regardless of size or industry and whether they sell online or in person. It begins with gathering and analyzing data and ends with acting on the information gathered in a way that is appropriate and effective.

source : shopify

Customer segmentation is used to predict in a vast demographic dataset, what part should be targeted by the company for a specific marketing goal. Therefore the machine learning techniques are at their best because the number of people in the data can be very high and the number of parameters for each person can also be very high and turn this into an high dimension problem.

Machine learning offers a solid solution to this kind of problem

Problem statement :

How can the client's company, acquire new customers more efficiently ?

This is a classification problem :

- For the unsupervised part the model takes the demographic data as input and output a corresponding cluster.
- For the supervised part the model takes also demographic data as input and output a class corresponding to the 'RESPONSE' column of the data

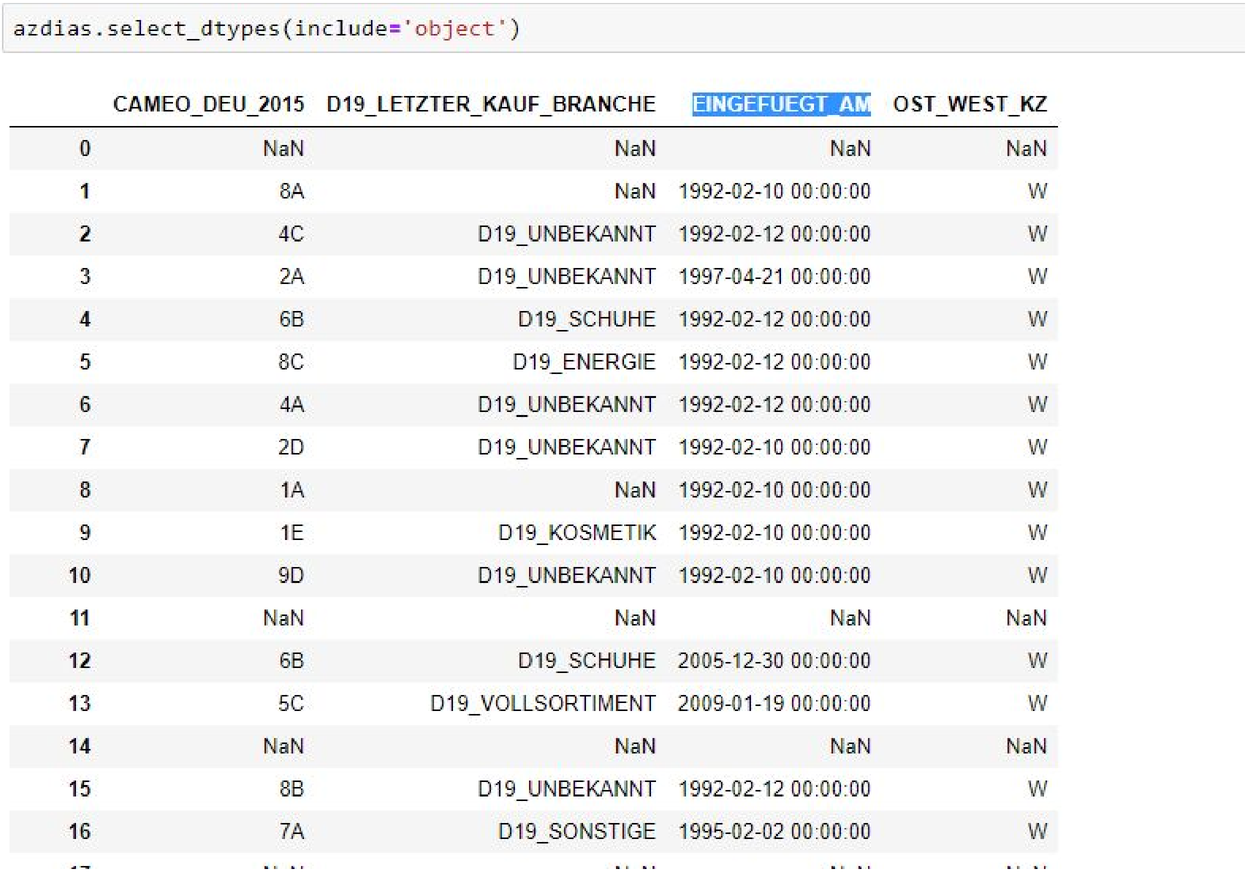
Datasets and inputs :

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).

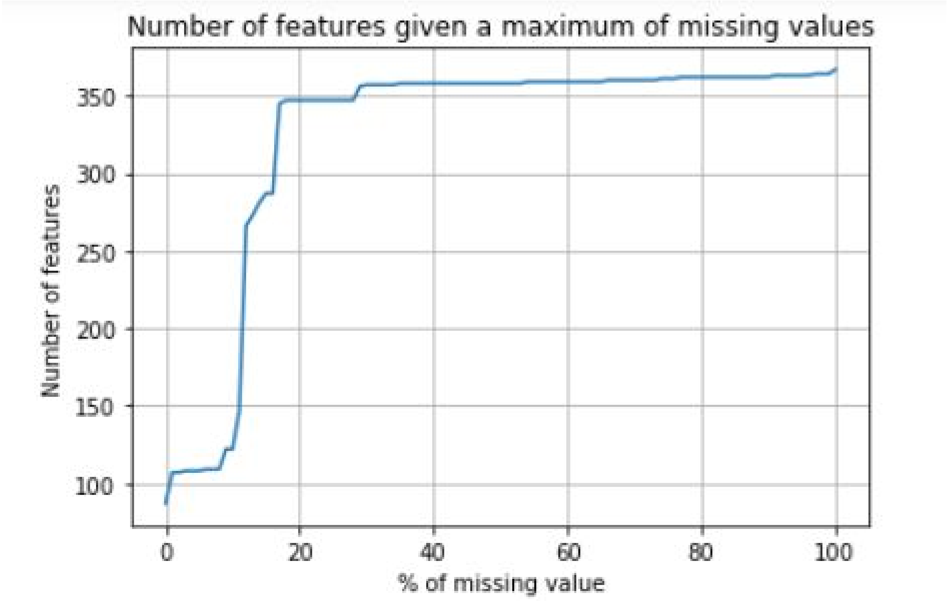
  Most of the columns have numerical values, a few have categorical values of type str, CAMEO_DEUG_2015 and CAMEO_INTL_2015 raise a warning because they have mixed type.

  EINGEFUEGT_AM represents dates in str format



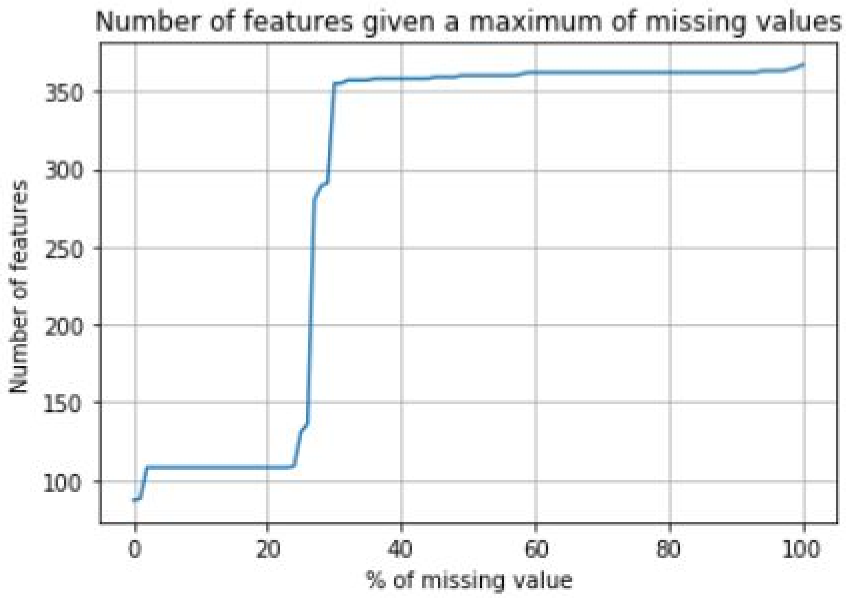
```
azdias.select_dtypes(include='object')
```

	CAMEO_DEU_2015	D19_LETZTER_KAUF_BRANCHE	EINGEFUEGT_AM	OST_WEST_KZ
0	NaN	NaN	NaN	NaN
1	8A	NaN	1992-02-10 00:00:00	W
2	4C	D19_UNBEKANNT	1992-02-12 00:00:00	W
3	2A	D19_UNBEKANNT	1997-04-21 00:00:00	W
4	6B	D19_SCHUHE	1992-02-12 00:00:00	W
5	8C	D19_ENERGIE	1992-02-12 00:00:00	W
6	4A	D19_UNBEKANNT	1992-02-12 00:00:00	W
7	2D	D19_UNBEKANNT	1992-02-10 00:00:00	W
8	1A	NaN	1992-02-10 00:00:00	W
9	1E	D19_KOSMETIK	1992-02-10 00:00:00	W
10	9D	D19_UNBEKANNT	1992-02-10 00:00:00	W
11	NaN	NaN	NaN	NaN
12	6B	D19_SCHUHE	2005-12-30 00:00:00	W
13	5C	D19_VOLLSORTIMENT	2009-01-19 00:00:00	W
14	NaN	NaN	NaN	NaN
15	8B	D19_UNBEKANNT	1992-02-12 00:00:00	W
16	7A	D19_SONSTIGE	1995-02-02 00:00:00	W

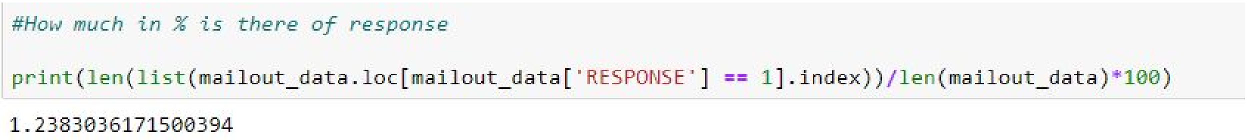


- Udacity_CUSTOMERS_052018.csv: Demographics data for customers of a mail-order company; 191 652 persons (rows) x 369 features (columns).

  Same thing as in Azdias except there more missing values and there are 3 feature in it that are not in azdias : ‘CUSTOMER_GROUP', 'ONLINE_PURCHASE', 'PRODUCT_GROUP'



- Udacity_MAILOUT_052018_TRAIN.csv: Demographics data for individuals who were targets of a marketing campaign; 42 982 persons (rows) x 367 (columns).



```
#How much in % is there of response

print(len(list(mailout_data.loc[mailout_data['RESPONSE'] == 1].index))/len(mailout_data)*100)
```

```
1.2383036171500394
```

  So the class is highly imbalanced as there is only 1.238 of response and everything else is then 0

- Udacity_MAILOUT_052018_TEST.csv: Demographics data for individuals who were targets of a marketing campaign; 42 833 persons (rows) x 366 (columns).

And two Excel spreadsheets :

- DIAS Information Levels - Attributes 2017.xlsx: is 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.

Solution statement :

As the data is real life data I will most likely need to clean and normalize it.

Then I will use PCA and then Kmeans on the general dataset, and then fit the customers data with this model.

Comparing the cluster from the 2 dataset will achieve the customer segmentation

Then for the supervised learning part, the goal is to use a model fitted with the mailout data to predict the 'RESPONSE' of the mailout_test dataset

I will need to evaluate the model to use, and then tune the hyperparameters to achieve the best result on Kaggle

Benchmark model :

As a benchmark I will use LightGBM as it is fast and scores very well in Kaggle's customers conversion that I've come across.

Evaluation metrics :

For the unsupervised part of the project we will use gap statistic analysis to find the number of cluster to choose for clustering

For the supervised part we will need to know the level of class imbalance to choose between the different options of evaluation metrics : precision, accuracy, recall, ...

If there is a big class imbalance I will avoid precision and accuracy and I will go for recall or ROC AUC for instance

Project design:

- Get to know the data :

  Here we will explore the data first, then I will clean it from unknown/missing values.

  Then we choose to keep only some of the columns that have less than a specific proportion of missing data.

  Then I encode all the non numerical columns remaining or reencode features that have mixed information in it.

  Then I replace all the missing data in each columns with the most frequent element of the column.

- Customer segmentation report :

  I use a MinMax Scaler to normalize the range of each columns.

  I perform a dimensionality reduction with PCA.

  I choose KMeans to cluster the dataset.

I use a gap statistic to choose the number of cluster of KMeans.

Then I use the general demographic dataset -> cleaning -> MinMax -> PCA -> KMeans.

Then I clean -> MinMax -> PCA the customers data and then transform it with the previously fitted KMeans.

Then I plot an histogram with side by side bins of the proportion for each cluster for the two dataset

Then the customer are segmented : with the previous plot we can see what cluster are more likely to hold future customers and which are not.

- Supervised learning model :

  First step is to clean and MinMaxScale the mailout data.

  Then check if I will use or not PCA.

  Then I will test the performance of : MLP, Logistic Regression, Random Forest, LGBM ,Gradient Boosting, XGB.

  Then to find best hyperparameters for the model I will perform a BayesSearchCV

  Then fit the model and predict the 'RESPONSE' probability of the mailout_test dataset

  Finally I will upload to kaggle the results to see my performance compared to others