Machine Learning Engineer Nanodegree

Capstone Project Proposal

Hafiz Muhammad Babar Sajjad
May 5th , 2020

Proposal

Domain Background

As the term *’data is the new oil’* [1] is getting more and more common nowadays therefore the companies are investing more in the fields of machine learning and data science as well. Many companies are benefitting from the use of machine learning based predictions to improve their customers base by conducting specially designed and focused marketing campaigns [2]. Similarly, the future customers will be predicted in this project for Arvato Financial Services as well.

Problem Statement

In this project ‘Create a Customer Segmentation Report for Arvato Financial Services’, the demographical data will be analysed for the customers of a mail-order company to investigate different patterns. The data is provided by the Arvato Financial Solutions. The idea will be to make different comparisons between the general population and the existing customers to increase the customer base in the future by conducting focused marketing campaigns.

There will be mainly four parts of it:

1. Get to know Data: data preprocessing will be performed here to investigate the given data and to reencode, reengineer the features which could be problematic in the later analysis.

2. Customer Segmentation Report: both general population and customer’s data will be used here for the application of unsupervised learning model to get the needed target population for the marketing campaigns.
3. Supervised Learning: labelled data will be used here to apply the supervised learning model to get the predictions on future customers.
4. Kaggle competition: in this part, the tuned supervised learning model will be applied to the test data set and the scores will be uploaded to the kaggle site to get the rating of the results.

Datasets and Inputs

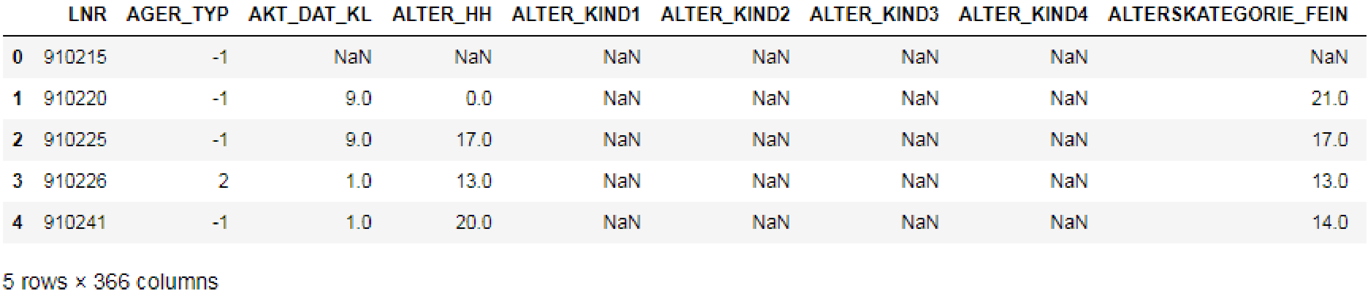
Following are the details for the given datasets:

- `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).
-

Additionally, there are two more files to describe the features:

- `DIAS Information Levels - Attributes 2017.xlsx`
- `DIAS Attributes - Values 2017.xlsx`

This is how the top 5 rows of demographics data `Udacity_AZDIAS_052018` looks like:

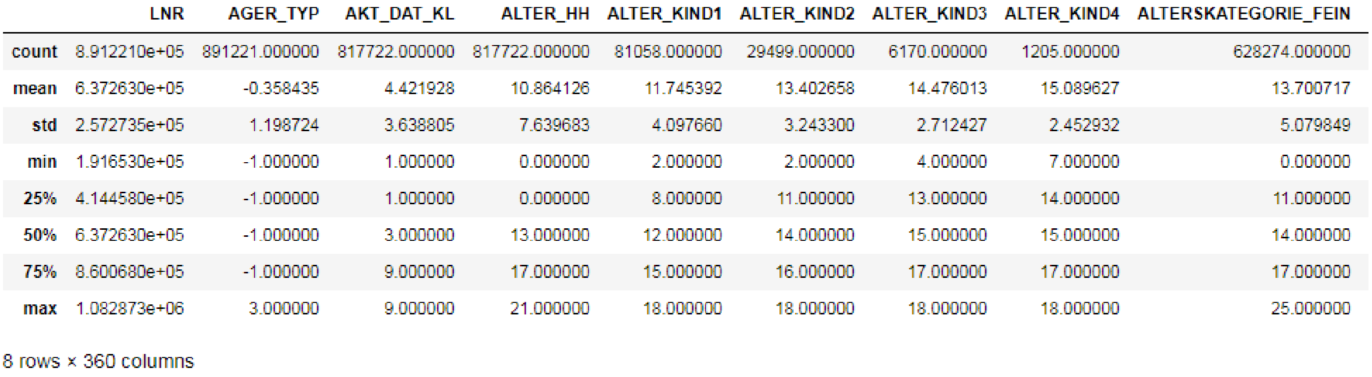


	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	ALTER_KIND1	ALTER_KIND2	ALTER_KIND3	ALTER_KIND4	ALTERSKATEGORIE_FEIN
0	910215	-1	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	910220	-1	9.0	0.0	NaN	NaN	NaN	NaN	21.0
2	910225	-1	9.0	17.0	NaN	NaN	NaN	NaN	17.0
3	910226	2	1.0	13.0	NaN	NaN	NaN	NaN	13.0
4	910241	-1	1.0	20.0	NaN	NaN	NaN	NaN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

5 rows × 366 columns

As you can see that there are many NaN values in the data. Also, there are some additional values in the data mentioned in the `DIAS Attributes - Values 2017` which should be replaced by NaNs.

Following is how the statistics of `Udacity_AZDIAS_052018` looks like:



	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	ALTER_KIND1	ALTER_KIND2	ALTER_KIND3	ALTER_KIND4	ALTERSKATEGORIE_FEIN
count	8.912210e+05	891221.000000	817722.000000	817722.000000	81058.000000	29499.000000	6170.000000	1205.000000	628274.000000
mean	6.372630e+05	-0.358435	4.421928	10.864126	11.745392	13.402658	14.476013	15.089627	13.700717
std	2.572735e+05	1.198724	3.638805	7.639683	4.097660	3.243300	2.712427	2.452932	5.079849
min	1.916530e+05	-1.000000	1.000000	0.000000	2.000000	2.000000	4.000000	7.000000	0.000000
25%	4.144580e+05	-1.000000	1.000000	0.000000	8.000000	11.000000	13.000000	14.000000	11.000000
50%	6.372630e+05	-1.000000	3.000000	13.000000	12.000000	14.000000	15.000000	15.000000	14.000000
75%	8.600680e+05	-1.000000	9.000000	17.000000	15.000000	16.000000	17.000000	17.000000	17.000000
max	1.082873e+06	3.000000	9.000000	21.000000	18.000000	18.000000	18.000000	18.000000	25.000000

8 rows × 360 columns

The information from `DIAS Information Levels - Attributes 2017` and `DIAS Attributes - Values 2017` will be used to get the list unknowns as well as the list for the features to reencode and reengineer.

Solution Statement

Analysis will be performed on the data during the preprocessing phase to replace the unknowns. Similarly, features will be reencoded, reengineered or dropped (if needed).

Then, imputation and scaling will be performed to apply the Principal Component Analysis (PCA) as there is a wide spread data and PCA helps in reducing the dimensionality of it [3]. Afterwards, kmeans clustering will be applied as it is very simple and popular technique to cluster the similar data points together and get to their underlying patterns [4].

Supervised learning model will be trained and tuned which will in the end be used for making the predictions on the test data. Following are the models which will be used in the project:

- RandomForestClassifier: It is a simple and fast algorithm to apply. Also suitable for handling large datasets with higher dimensions.[5]
- GradientBoostingRegressor: It belongs to the ensemble decision tree regressor models and can be used for the classification problems[6]. Also, it is one of the best algorithms when it comes to building the predictive models.

- XGBRegressor: It is a high performing and fast algorithm for the classification problems - and it is basically the fastest implementation of the Gradient boosting algorithm[7].

GridSearchCV will be used for hypertuning the models and to get the optimal parameters.

Benchmark Model

RandomForestClassifier will be used as a benchmark model as it is simple to use and is suitable for higher dimension classification problems. The other above mentioned models will also be tried, tuned and compared.

Evaluation Metrics

AUC-ROC curve will be used in the GridSearchCV to evaluate the results as it is mentioned on the Kaggle site.

Project Design

Following will be the steps in the project design :

- Data preprocessing
- Replacing unknowns with nans.
- Column and row wise nan calculations
- Features reencoding and reengineering
- Dropping unnecessary data.
- Performing get_dummies, imputation and scalling
- Applying PCA and Kmeans
- Applying supervised learning models
- Make predictions on test data
- Submit results to Kaggle

References

[1] https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data
[2] https://digitalmarketinginstitute.com/en-eu/blog/7-ways-machine-learning-can-enhance-marketing
[3] https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202
[4] https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1
[5] https://www.quora.com/What-are-the-advantages-and-disadvantages-for-a-random-forest-algorithm
[6] https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/
[7] https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/