Machine Learning Engineer Nanodegree

Capstone Project Report

Hafiz Muhammad Babar Sajjad
May 7th , 2020

Domain Background and Problem Statement

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 are predicted in this project for Arvato Financial Services as well.

In this project 'Create a Customer Segmentation Report for Arvato Financial Services', the demographical data is 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:

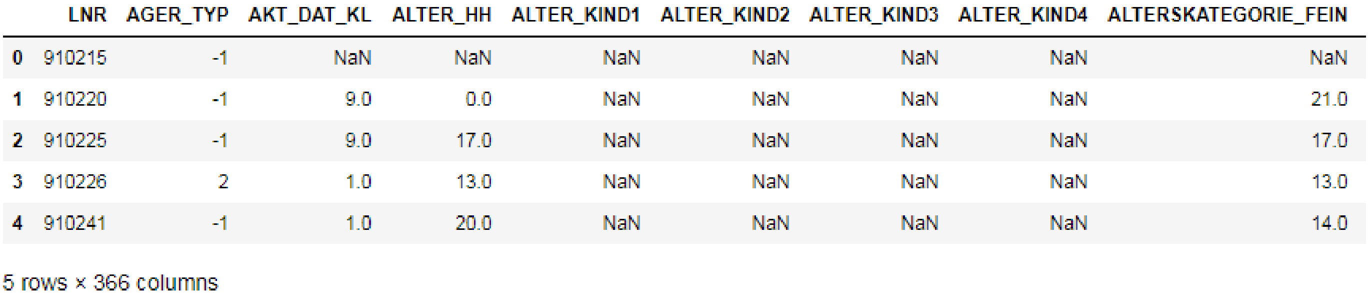
- `DIAS Information Levels - Attributes 2017.xlsx`

- `DIAS Attributes - Values 2017.xlsx`

Get to know Data

In the real life projects, understanding and cleaning the data is the most important phase. For this project, the given data had all sort of issues present in it therefore it required thorough investigation.

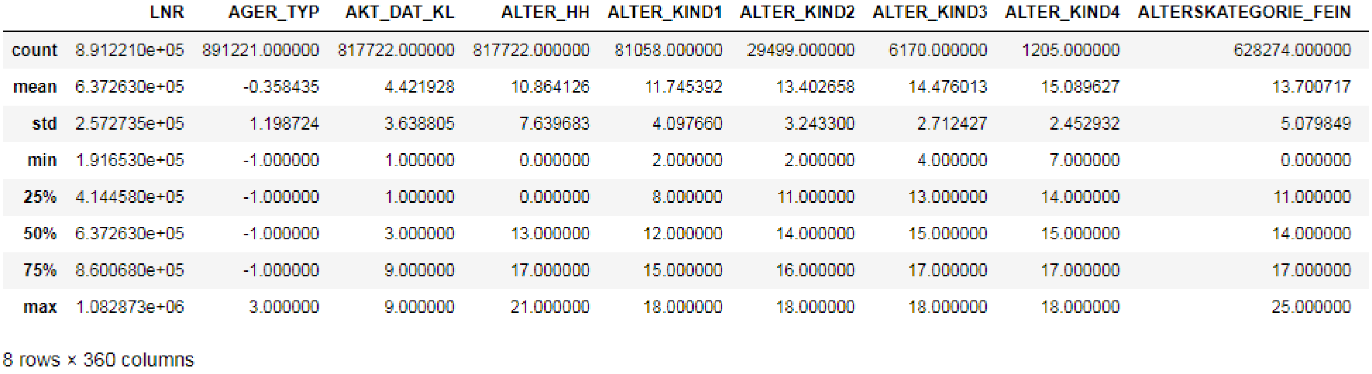
This is how the top 5 rows of demographics data `Udacity_AZDIAS_052018` looked like:



	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

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



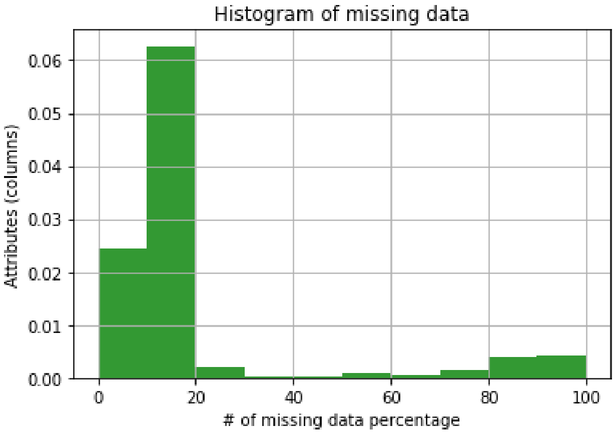
	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	ALTER_KIND1	ALTER_KIND2	ALTER_KIND3	ALTER_KIND4	ALTERSKATEGORIE_FEIN
count	8.912210e+05	891221.000000	817722.000000	817722.000000	81058.000000	29499.000000	6170.000000	1205.000000	628274.000000
mean	6.372630e+05	-0.358435	4.421928	10.864126	11.745392	13.402658	14.476013	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

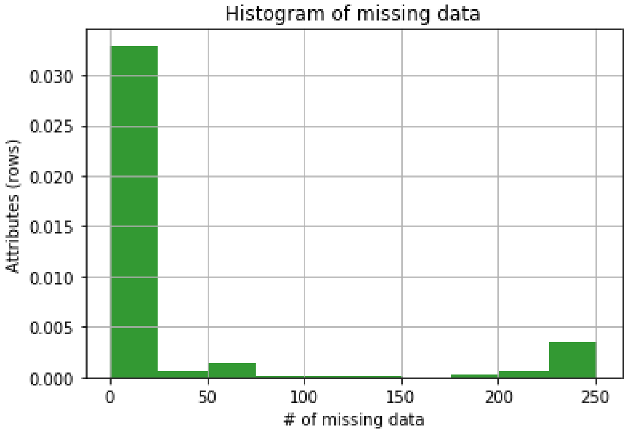
8 rows × 360 columns

There were many NaN values in the data. The information from DIAS Information Levels - Attributes 2017 and DIAS Attributes - Values 2017 was used to get the list of unknowns as well as the list for the features to re-encode and re-engineer. Also, the description or explanation was not given for all the features.

A combined feature information file was created out of the two given files explaining the nans for each attribute and their respective type i.e. categorical, ordinal mixed etc.

Following two histograms are showing the percentage and numbers of NaNs in columns and rows.





There were 41 columns with more than 30 percent of nans and there were 154916 rows with more than 25 nans – which were removed from the data.

The next part was about re-encoding and re-engineering the features and in the end of this part, the unnecessary features were also removed from the main data set.

Special handling was needed for the variable types: categorical and mixed. Decisions were made on whether to keep, drop, or re-encode each. Then, in the last part, a new data frame with only the selected and engineered columns was created. For binary data, nothing was done whereas the non-numeric two state data was replaced by numeric binary data. For multi-level data, get_dummies was used.

In summary:

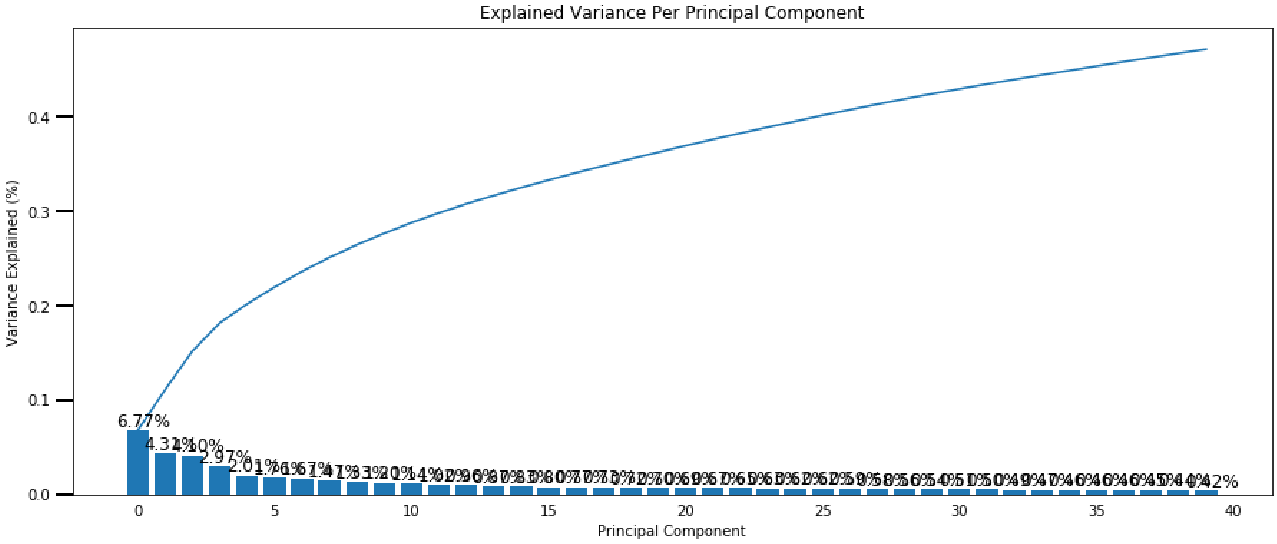
- OST_WEST_KZ reencoded with 'W' = 0, else 1
- PRAEGENDE_JUGENDJAHRE_D and PRAEGENDE_JUGENDJAHRE_M created from PRAEGENDE_JUGENDJAHRE based on 'decade' and 'movement' info.
- NEIGHBOUR_RURAL is created from WOHNLAGE based on rural and neighbourhood info.
- CAMEO_INTL_2015_W and CAMEO_INTL_2015_L created from CAMEO_INTL_2015 based on wealth and life info.
- PLZ8_BAUMAX_Family and PLZ8_BAUMAX_Business created from PLZ8_BAUMAX based on the info family and business info.
- Remaining 'mixed' types are removed from the dataset because of many 0s - also the orignal features were removed as the encoded and engineered feature were added.

- VERDICHTUNGSRAUM, ANZ_KINDER, CAMEO_DEU_2015 also removed because of so many multi levels.
- ANZ_TITEL, ANZ_HH_TITEL dropped because of many 0s.
- LP_LEBENSPHASE_FEIN, LP_LEBENSPHASE_GROB dropped because of complex data.

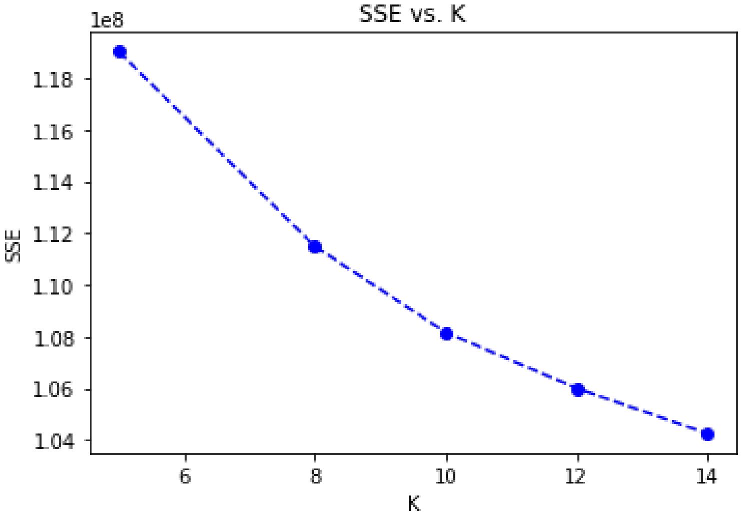
Imputation was used to transform the nans to the most_frequent values and then a cleaning function was created to simplify this process for the other files.

Customer Segmentation Report

Principal Component Analysis (PCA) was used as there was a wide spread data and it helps in reducing the dimensionality of it [3]. There was a memory issue during the execution therefore PCA was replaced by IPCA. Scaling was performed on the data as well before applying the IPCA. The number of n_components was set to 40 as the most of the varience was already explained by the first 15 componets.

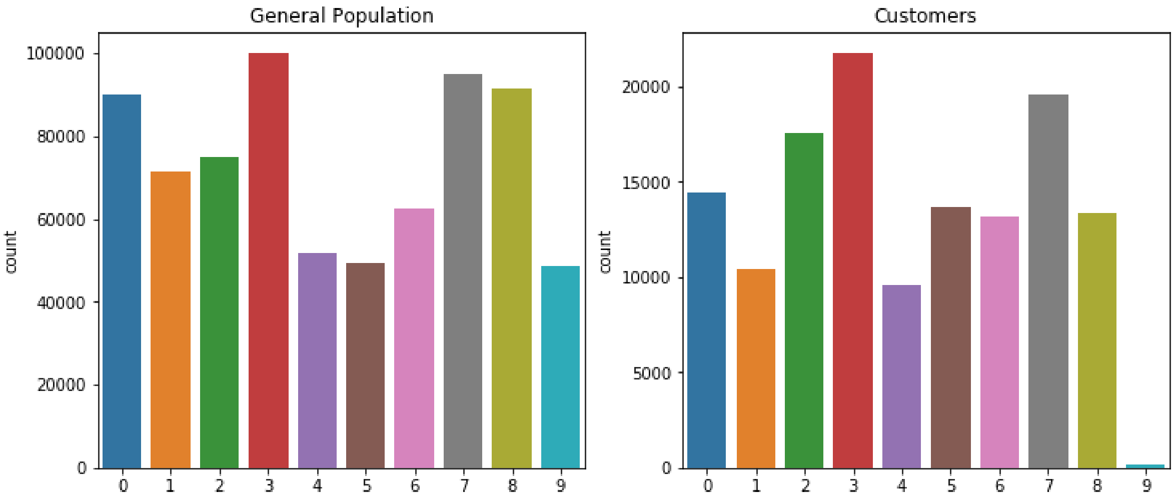


Afterwards, kmeans clustering was applied as it is very simple and popular technique to cluster the similar data points together and get to their underlying patterns [4]. Elbow method was used to find the number of clusters which was then set to 10.



Comparison with Customers data

Customers data was cleaned, scaled and then PCA and kmeans clustering was applied. Also, a function was created to map the weights for the first principal component to corresponding feature names and then print the linked values, sorted by weight.



In the end of this section, the Customer Data was compared to the Demographics data. As can be seen from the graphs above that cluster 9 is under-represented for the customers. The main segments of the population which are underrepresented were e.g.:

- Number of household in the buiding (ANZ_HAUSHALTE_AKTIV -> -4.74)
- Number of adults in household (ANZ_PERSONEN -> -0.96)

Similarly, cluster 5 looks like over-represented and the main segments of the population were e.g.:

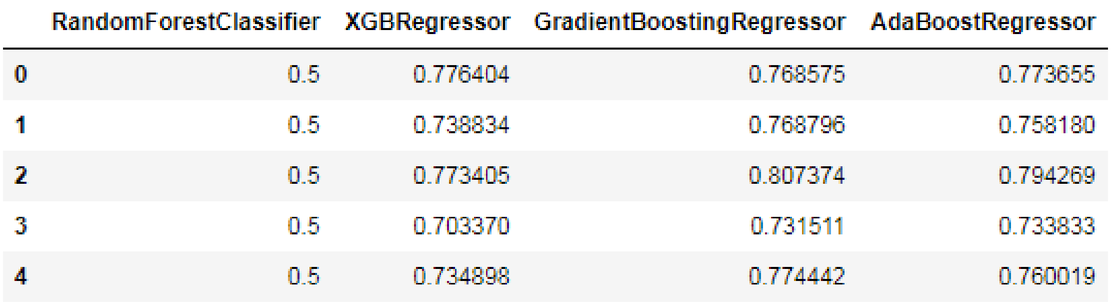
- Number of household in the buiding (ANZ_HAUSHALTE_AKTIV -> 2.97)
- Number of adults in household (ANZ_PERSONEN -> 1.69)

Supervised Learning Model

After finding out which parts of the population are more likely to become the future customers of the company, supervised learning models were trained, tuned and compared to get the best one for making the predictions. Following models were used in the project by keeping in mind the problem to solve – which is a classification one but with big data set and dimensions:

- 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].
- AdaboostRegressor: Also belongs to the ensemble family and gives good classification results with much less tweaking of parameters or settings.

GridSearchCV was used be used for hypertuning the models and to get the optimal parameters and the Stratified Kfold was used for cross-validation. AUC-ROC curve was used in the GridSearchCV to evaluate the results as it is more suitable for this problem and it is also mentioned on the Kaggle site. Following are the results with the basic configurations.



	RandomForestClassifier	XGBRegressor	GradientBoostingRegressor	AdaBoostRegressor
0	0.5	0.776404	0.768575	0.773655
1	0.5	0.738834	0.768796	0.758180
2	0.5	0.773405	0.807374	0.794269
3	0.5	0.703370	0.731511	0.733833
4	0.5	0.734898	0.774442	0.760019

Benchmark Model

RandomForestClassifier was used as a benchmark model as it is simple to use and is suitable for higher dimension classification problems. The initial results were not that good but when it was used with the GridSearchCV to tune the n_estimators and max_depth parameters – it gave a much better score of 0.747.

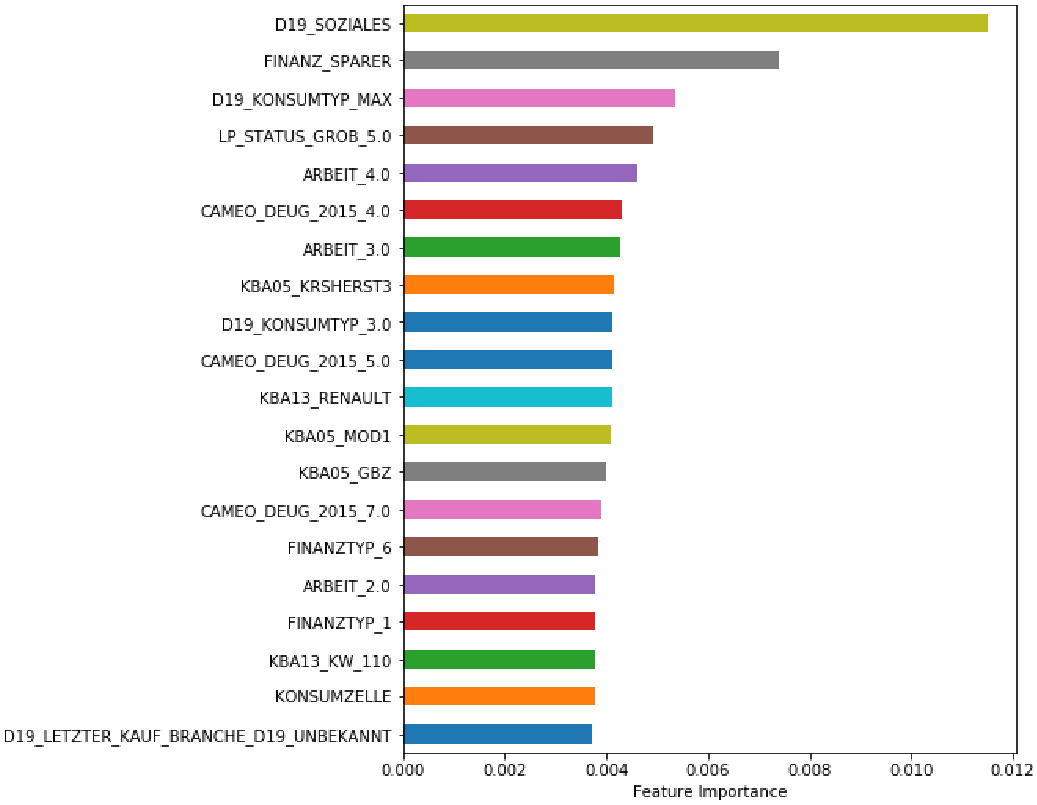
Other Models:

GradientBoostingegressor shown good results with the basic configuration. It was then used with the GridSearchCV to tune the min_samples_split and max_depth parameters – it gave a much better score of 0.784 but this hypertuning exercise took about 5 hours to finish therefore I decide to move on with the other models.

XGBoostRegressor was used with the configuration of objective = 'binary:logistic' and scale _pos_weight = 1 as it was being used for the classification problem and additionally there was the issue of high class imbalance. Initially, only the parameters of ‘max_depth’' and min_child_weight were used with GridSearchCV for tuning and the resulting score was 0.762

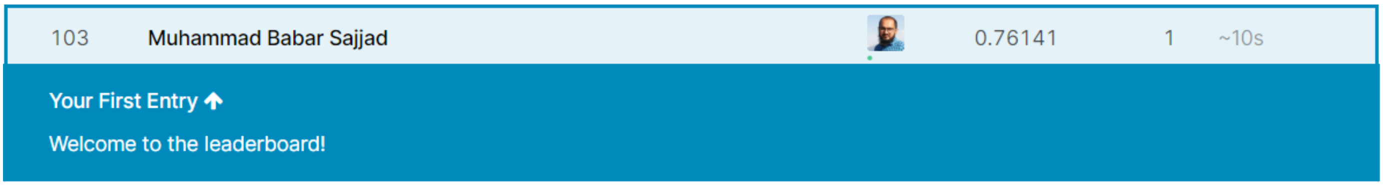
It was again tuned with the additional parameters of ‘learning_rate’ and ‘n_estimators’. The resulting score improved to 0.78. It also gave an impression of further reducing the learning_rate and increasing the n_estimators. Therefore, Stratified Kfold was also used here for cross-validation and all other best parameters were retained. The final average score was ~0.78 with the best score of 0.806 and hence it was choosen as a final model to predict against the test data set.

Following are the feature importances from the XGBoostRegressor. It can been seen the most important feature is D19_SOZIALES. Although the description is unfortunately not provided for this feature but if it is somehow related to the social insurance - then that can for sure tell a lot about the people.



Kaggle competition

Also, I realized at this stage that the cleaning function was dropping the rows for mailout_train and mailout_test data as well as it was designed for the azdias dataset. I modified it, but due to the time constraint I could not tune and analyze the models again. I only trained the best performing model with Stratified Kfold cross-validation – the resulting average score on the mailout_train was 0.74091 and on the Kaggle mailout_test was 0.76141 . I will continue to work on it together with other parts mentioned in the Future Work.



103 Muhammad Babar Sajjad 0.76141 1 ~10s

Your First Entry ↑

Welcome to the leaderboard!

Future Work

- Get to know data: In this part, features can further be explored and then accordingly re-encoded and re-engineered.
- Supervise Learning Model:
    - Adaboost model can also the used with the GridSearchCV to findout the best scores with it as well.
    - Learning Curve can also be tried to see how helpful it is.
    - XGBoost model can further be explored as it is one of the best performing models nowadays.

Conclusion

In this real world project, the valuable data from Arvato Financial Solutions was used to analyze the general population demographical information to increase the customer base by using the unsupervised and supervised machine learning algorithms.

The most challenging but interesting part was about preprocessing the data and getting familiar with it. There were different data types, differently encoded missing and unknown values. Also, there were many features without any description. It was a great learning experience in getting familiar with such type of data and to preprocess it in a way to get the most information out of it.

By applying the PCA and Kmeans unsupervised learning algorithms to the segment of general population, it was possible to predict the future customers of the company as well. The clusters could further be interpreted to get more information out of them.

Also, by applying the supervised learning algorithms, it was possible to make predictions at the person level that whether the person would become the future customer or not.

Finally, the best performing supervised learning algorithm was used to make predictions on the test data.

In the end, I would like to thank Udacity and Bertelsmann Arvato Analytics for organizing this real world project.

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/