Machine Learning Engineer Nanodegree

Capstone Project Report

Joe Abraham

September 26, 2020

Definition

The objective of this project is to analyze demographics data of customers of a mail-order sales company in Germany that sells organic products and compare that to demographics data of the general population of Germany. The end goal is to be able to predict, based on demographics data, which individuals from the general population should be targeted in the mail-order campaign. Both unsupervised and supervised learning techniques will be used. Unsupervised learning will be used to help identify segments of the general population of Germany that best matches the existing customer base of the company. A supervised learning prediction model will be developed to predict the likelihood of whether or not an individual of the general population will become a customer. The dataset was provided by a real business - Bertelsmann Arvato Analytics and represents a real-life data science task.

The whole project is divided into 4 parts:

- Part 0 : Know the data, where you try to understand the data
- Part 1 : Customer Segmentation (Unsupervised Learning)
- Part 2 : Supervised Learning
- Part 3 : Kaggle competition

Domain Background

About Arvato Financial Solutions

Arvato Financial Solutions offers all financial services related to payments and cash flow. They provide professional financial services to renowned international brands as well as respected local businesses — allowing their customers to leave their

credit management to a professional, and they can focus on what matters most for their business.

Their services center around cash flow in all segments of the customer lifecycle: from identity, fraud and credit risk management, to payment and financing services and debt collection.

Source : https://www.linkedin.com/company/arvato-financial-solutions/about

Project Background

In this project, Arvato requires a customer segmentation report and wants to predict their potential customers for a mail campaign who will mostly reply back.

Problem Statement

The Arvato Financial Solutions wants to find the potential customers, additionally they want to know more about the customers and their background. For this reason, analysis of demographics data for customers is done against demographics information for the general population.

1. How to identify parts of the general population to be ideal customers?
2. A method to identify individuals who are likely to be a customer ?

Datasets and Inputs

There are four data files associated with this project:

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).

Additional support files available :

1. DIAS Information Levels - Attributes 2017.xlsx is a top-level list of attributes and descriptions, organized by informational category.
2. DIAS Attributes - Values 2017.xlsx is a detailed mapping of data values for each feature in alphabetical order.

Due to privacy concerns - Datasets are protected and are not publicly available.

Benchmark Model

Supervised Learning, Logistic regression is used as a benchmark model.

Evaluation Metrics

Performance measurement is an essential task. Since our problem is a classification problem, we can use AUC - ROC Curve to better understand our model. And how our model works. AUC - ROC curve is a performance measurement for classification problems at various thresholds settings. A ROC, or receiver operating characteristic, is a graphic used to plot the true positive rate (TPR, proportion of actual customers that are labeled as so) against the false positive rate (FPR, proportion of non-customers labeled as customers).

Source :
https://www.kaggle.com/c/udacity-arvato-identify-customers/overview/evaluation

Solution Outline

Part 0 - Knowing the Data

Data preprocessing - Inspect the data and necessary data cleaning needs to be done.

Part 1 - Unsupervised learning

Clustering needs to be done with the demographic data provided in Azdias (general population) and customers. For that PCA and K-Means clustering will be the best method and clustering needs to be done with the best model.

Part 2 - Supervised Learning

Here the supervised learning algorithms have to be applied on Udacity_MAILOUT_052018_TRAIN.csv and the best performing model will be tuned for best results.

Part 3 : Kaggle competition

Predictions are made with the developed model for Udacity_MAILOUT_052018_TRAIN.csv and submitted on the Kaggle

Solution

Part 0 - Knowing the Data

Data Exploration, Preprocessing and Cleaning

Data Exploration

How a sample of general population dataset looks like

```
In [41]: display(azdias.head())
```

	Unnamed: 0	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	ALTER_KIND1	ALTER_KIND2	ALTER_KIND3	ALTER_KIND4	ALTERSKATEGORIE_FEIN	ANZ_HA
0	0	910215	-1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	1	910220	-1	9.0	0.0	NaN	NaN	NaN	NaN	21.0	11.0
2	2	910225	-1	9.0	17.0	NaN	NaN	NaN	NaN	17.0	10.0
3	3	910226	2	1.0	13.0	NaN	NaN	NaN	NaN	13.0	1.0
4	4	910241	-1	1.0	20.0	NaN	NaN	NaN	NaN	14.0	3.0

5 rows × 367 columns

How a sample of customer dataset looks like

```
In [41]: display(azdias.head())
```

	Unnamed: 0	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	ALTER_KIND1	ALTER_KIND2	ALTER_KIND3	ALTER_KIND4	ALTERSKATEGORIE_FEIN	ANZ_HA
0	0	910215	-1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	1	910220	-1	9.0	0.0	NaN	NaN	NaN	NaN	21.0	11.0
2	2	910225	-1	9.0	17.0	NaN	NaN	NaN	NaN	17.0	10.0
3	3	910226	2	1.0	13.0	NaN	NaN	NaN	NaN	13.0	1.0
4	4	910241	-1	1.0	20.0	NaN	NaN	NaN	NaN	14.0	3.0

5 rows × 367 columns

How unknown values are represented in each column

Attribute	Value	Meaning
AGER_TYP	-1	unknown
ALTERSKATEGORIE_GROB	-1, 0	unknown
ALTER_HH	0	unknown / no main age detectable
ANREDE_KZ	-1, 0	unknown
BALLRAUM	-1	unknown
BIP_FLAG	-1	unknown
CAMEO_DEUG_2015	-1	unknown
CAMEO_DEUINTL_2015	-1	unknown
CJT_GESAMTTYP	0	unknown
D19_KK_KUNDENTYP	-1	unknown
EWDICHTE	-1	unknown
FINANZTYP	-1	unknown
FINANZ_ANLEGER	-1	unknown
FINANZ_HAUSBAUER	-1	unknown
FINANZ_MINIMALIST	-1	unknown
FINANZ_SPARER	-1	unknown
FINANZ_UNAUFFAELLIGER	-1	unknown
FINANZ_VORSORGER	-1	unknown
GEBAEUDETYP	-1, 0	unknown
GEOSCORE_KLS7	-1, 0	unknown
HAUSHALTSSTRUKTUR	-1, 0	unknown
HEALTH_TYP	-1	unknown
HH_EINKOMMEN_SCORE	-1, 0	unknown
INNENSTADT	-1	unknown
KBA05_ALTER1	-1, 9	unknown
KBA05_ALTER2	-1, 9	unknown
KBA05_ALTER3	-1, 9	unknown
KBA05_ALTER4	-1, 9	unknown
KBA05_ANHANG	-1, 9	unknown
KBA05_ANTG1	-1	unknown
KBA05_ANTG2	-1	unknown
KBA05_ANTG3	-1	unknown
KBA05_ANTG4	-1	unknown
	-1, 9	unknown
KBA05_BAUMAX	-1, 0	unknown
KBA05_CCM1	-1, 9	unknown

Data Pre-Processing and Cleaning

After data exploration I came up with a function to clean the dataset, I have developed a function to clean the dataset. Refer the function data_cleaning in the notebook.

I have done the following steps in pre-processing and cleaning

- Took the help of excel sheet to exactly know how unknown data is represented
- Filled the unknown data with NaNs
- Replace birth year with NaN, if it is zero.
- Dropped the columns, if most of the values are NaNs(only for unsupervised learning)
- Replaced NaN with most frequent value in that particular column
- Performed feature scaling
- If the cleaning is done for clustering, Remove the columns if most of the values are NaNs

Part 1 - Unsupervised learning - Customer Segmentation Report

The steps done for Unsupervised learning are

1. Dimension Reduction - Done by doing PCA
2. Clustering - K means clustering

Dimension Reduction

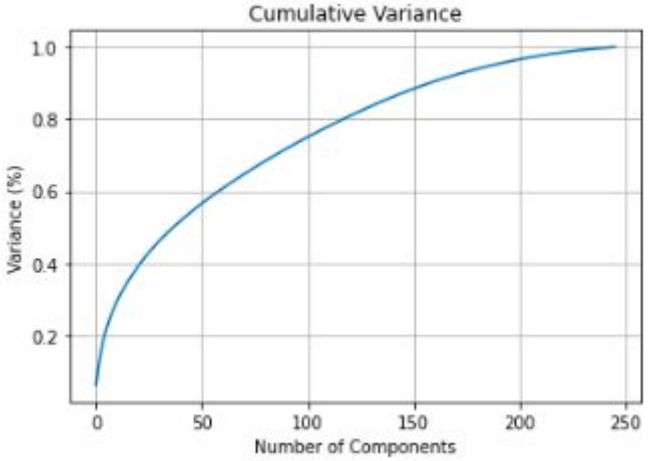
Dimension reduction is done by help of PCA. Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. Because smaller data sets are easier to explore and visualize and make analyzing data much easier and faster for machine learning algorithms without extraneous variables to process.

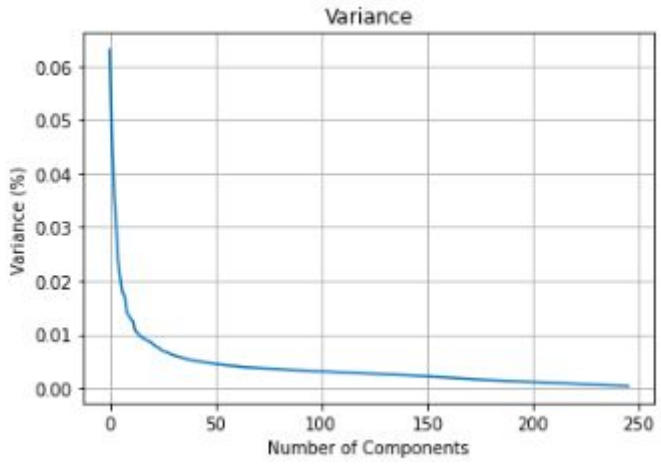
So to sum up, the idea of PCA is simple — reduce the number of variables of a data set, while preserving as much information as possible

Source : https://builtin.com/data-science/step-step-explanation-principal-component-analysis

The cumulative variance of the components is illustrated in the below graph



And the percentage of variance of components is illustrated in the below graph

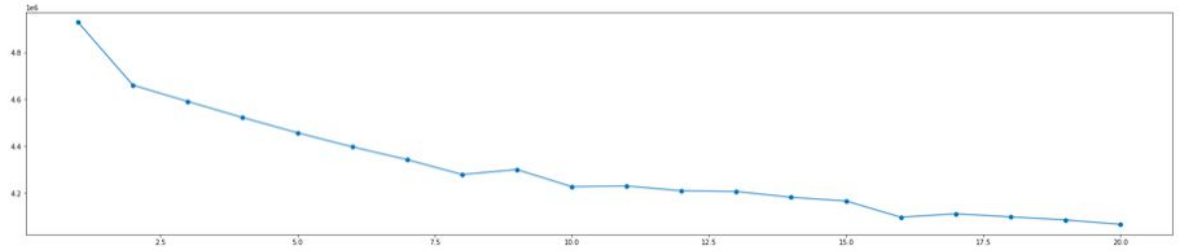


By close inspection we can notice that above 200 the variance is almost zero. Therefore only 200 Components are needed to plot the clusters. And it will give the best result.

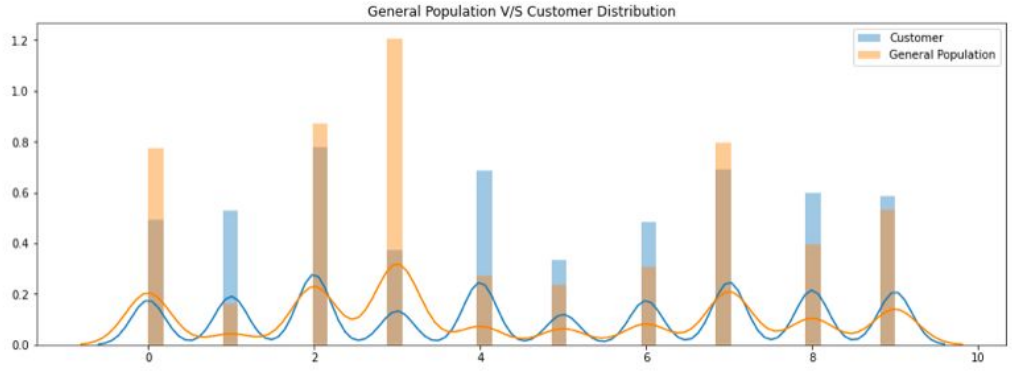
Clustering

After dimension reduction, clustering needs to be done, Here I am using K-means clustering technique. K-Means Clustering is an unsupervised machine learning algorithm. K-Means attempts to classify data without having first been trained with labeled data. Once the algorithm has been run and the groups are defined, any new data can be easily assigned to the most relevant group.

The data here has multiple dimensions making it difficult to visual. A way of determining this mathematically. We graph the relationship between the number of clusters and Within Cluster Sum of Squares (WCSS) then we select the number of clusters where the change in WCSS begins to level off simply called elbow method, and the graph is called elbow graph and it is illustrated below for our dataset. The best value of elbow was 10 from the graph



Final Conclusion in K-Means clustering



The above graphs shows that K-Means clustering with K = 10 helps to find customer from general public, it has very small difference, compared to others

Conclusion on Unsupervised Learning.

200 components with a K Value of 10 gave me the best result for clustering

Part 2 - Supervised Learning

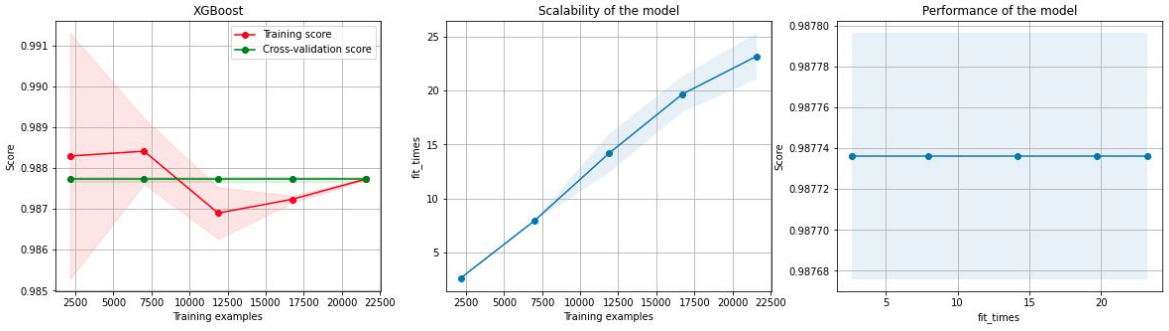
First and foremost step Data preprocessing

4 Different algorithms used to study about the supervised machine learning algorithms

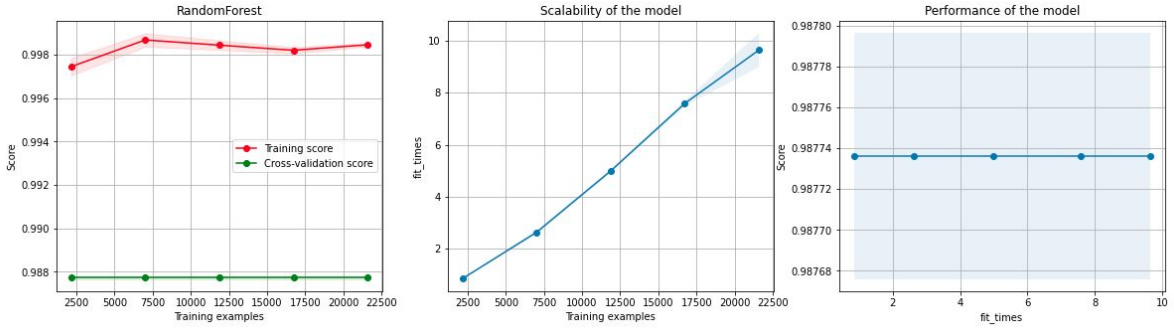
1. XGBoost,
2. Random Forest,
3. AdaBoost and
4. Decision Tree

Logistic regression is done and AUC score is
plot_learning_curve is taken from sklearn's official site. And plotted to know how the different algorithms work on the dataset.

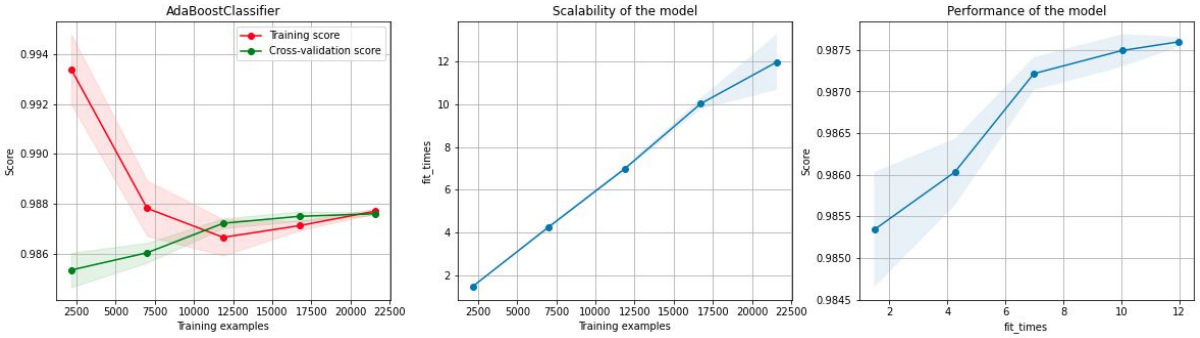
Learning Curve of different algorithms look like
XGBoost



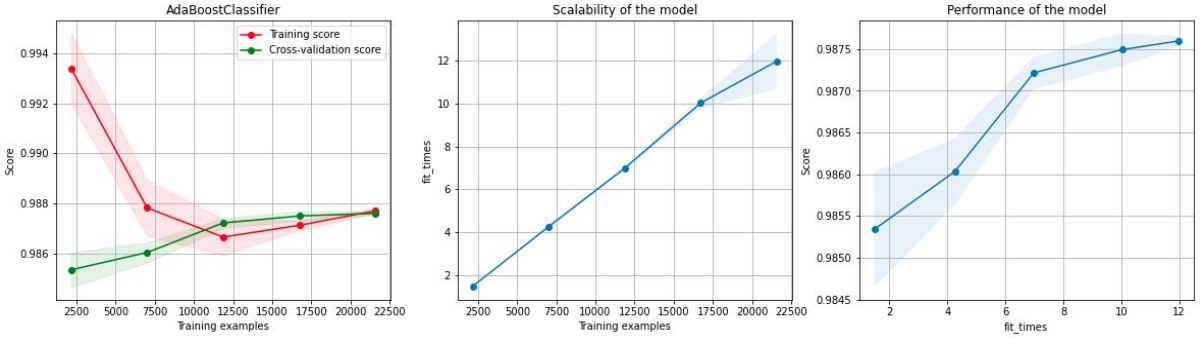
Random Forest



AdaBoost



Decision Tree



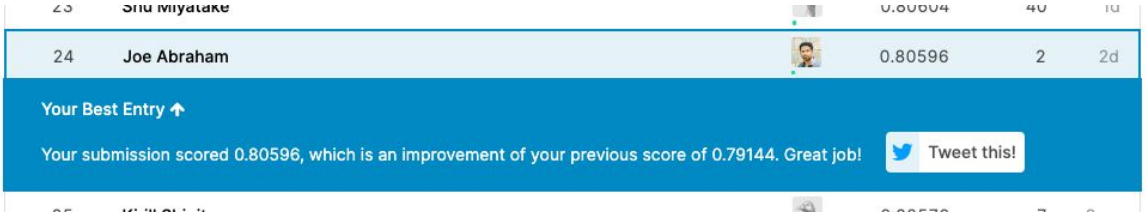
And I found that XGBoost is performing well.

Hyperparameter Tuning

In machine learning, hyperparameter optimization or tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. The mean absolute error of our base model was 0.2. After hyperparameter tuning, the mean absolute error is brought down to 0.15 and the cross validation score came out to be 0.9876169638016635

Part 3 : Kaggle competition

The model was run on a provided dataset and then submitted on Kaggle and my model performed with a score 0.80596. And I grabbed 24th place in leadership board as on 20-09-2020

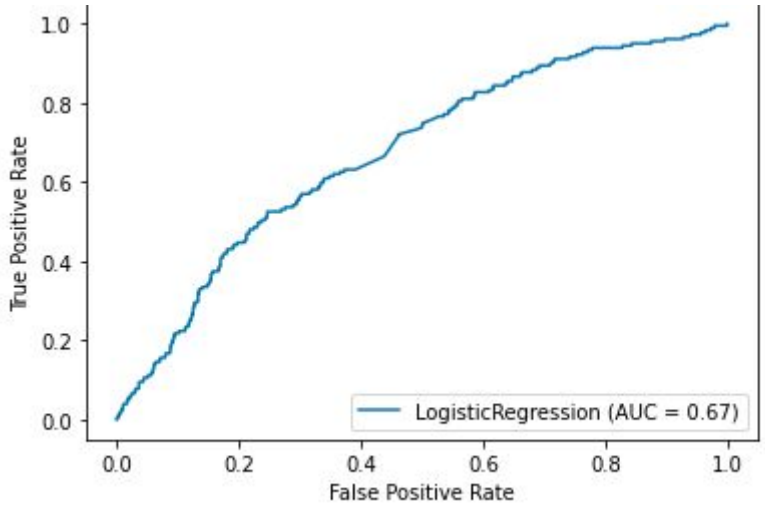


Evaluation Metrics

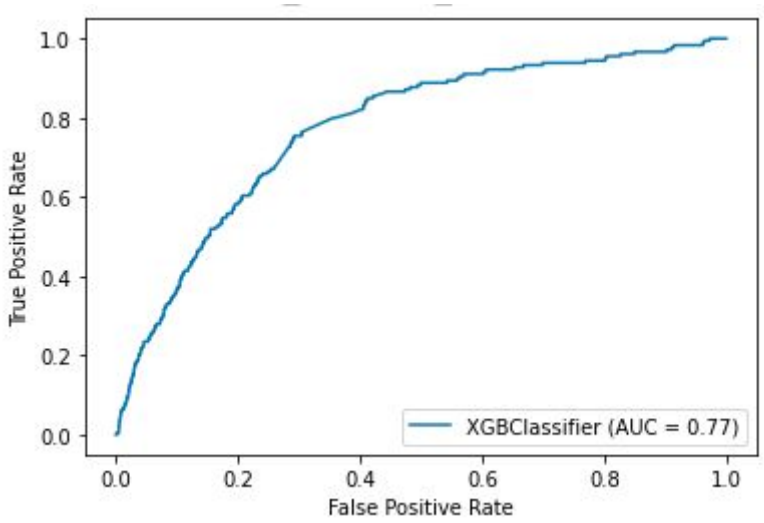
Performance measurement is an essential task. Since our problem is a classification problem, we can use AUC - ROC Curve to better understand our model. And how our model works. AUC - ROC curve is a performance measurement for classification problems at various thresholds settings. A ROC, or receiver operating characteristic, is a graphic used to plot the true positive rate (TPR, proportion of actual customers that are labeled as so) against the false positive rate (FPR, proportion of non-customers labeled as customers).

Source :
https://www.kaggle.com/c/udacity-arvato-identify-customers/overview/evaluation

AUC curve of base model is as below



AUC curve of final model


True Positive Rate
False Positive Rate
XGBClassifier (AUC = 0.77)
0.0
0.2
0.4
0.6
0.8
1.0