



---

# MACHINE LEARNING ENGINEER NANODEGREE

---

Capstone Project Report

Customer Segmentation - Arvato Financial Solutions



APRIL 6, 2020  
SANTOSH KANNAN

# Contents

<b>1. Introduction.....</b>	<b>2</b>
<b>1.1. Domain Background.....</b>	<b>2</b>
<b>1.2. Problem Statement.....</b>	<b>2</b>
<b>1.3. Datasets and Inputs: .....</b>	<b>2</b>
<b>1.4. Evaluation Metrics .....</b>	<b>3</b>
<b>2. Data Profiling.....</b>	<b>4</b>
<b>2.1. Data Exploration and Pre-processing .....</b>	<b>4</b>
2.1.1. Different datatype information in the same column .....	4
2.1.2. Addressing the issues with LP columns .....	4
2.1.3. Resolving Unknown Values .....	4
2.1.4. Mutual columns .....	4
2.1.5. Encoding Features .....	4
2.1.6. Missing Values.....	5
2.1.7. Imputing Missing Values .....	5
2.1.8. Feature Scaling .....	6
<b>3. Algorithms and Techniques .....</b>	<b>6</b>
<b>3.1. Customer Segmentation .....</b>	<b>6</b>
3.1.1. Dimensionality Reduction .....	6
3.1.2. Clustering .....	6
<b>3.2. Customer Acquisition.....</b>	<b>8</b>
3.1.1. Benchmark Model .....	8
3.1.2. Model Performance .....	8
3.1.3. Hyperparameter Tuning .....	8
<b>4. Conclusion and Results .....</b>	<b>9</b>
<b>5. Kaggle Submission .....</b>	<b>10</b>
<b>6. Future Scope.....</b>	<b>10</b>

# 1. Introduction

## 1.1. Domain Background

Arvato is a services company that develops and provides Supply Chain Management (SCM) solutions, Financial Services and IT services to clients all over the world. Key highlights include development of innovative solutions with focus on automation and data analytics. List of clients for Arvato range from insurance companies, e-commerce, energy providers to internet providers [1]. Arvato is a venture of the Bertelsmann group, which is giant in the media and education industry.

The business model for Arvato revolves around helping clients with data analytics to assist with key business decisions. This includes uncovering hidden patterns, identifying customer behaviour from the raw data. This strategy for customer centric clients utilises various aspects of Data Science and Machine Learning.

In this project, Arvato is assisting a Mail-Order company in Germany, which sells organic products, to understand its customers. The ask is to identify the possible future customers in a deterministic manner. On the basis of existing customer data across Germany, I propose a Machine Learning model which can identify probable customers.

## 1.2. Problem Statement

On the basis of existing demographics data for customers of a mail-order company in Germany, analyse and identify individuals who are most likely to convert into customers for the company.

The proposed solution will utilise an unsupervised model for the purpose of identifying customer segments in the dataset. This will be followed up by a supervised model which uses demographic information of target customers for the advertising campaign and predicts the individuals who are likely to convert into customers.

## 1.3. Datasets and Inputs:

The datasets are provided by Arvato for Udacity Machine learning Nanodegree program's project. The dataset consists of demographic data of customers in Germany. This comprises of 4 datasets and 2 metadata files associated with them.

Data File	Description
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).
DIAS Information Levels - Attributes 2017.xlsx	Top-level list of attributes and descriptions, organized by informational category
DIAS Attributes - Values 2017.xlsx	Detailed mapping of data values for each feature in alphabetical order.

## 1.4. Evaluation Metrics

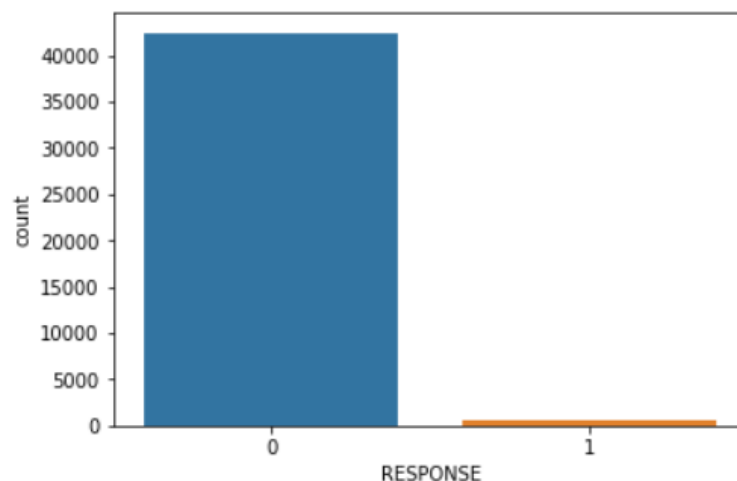
For the purpose of evaluating the model, we split the model into two phases:

### Phase 1: Customer classification with unsupervised algorithms

This phase implements a clustering algorithm on the customer dataset to understand the patterns in the dataset and uncover information. Before implementing the clustering algorithm, it is essential to reduce the dimensions in the dataset as the number of dimensions in the original dataset is relatively high. We thus use the technique known as Principal component Analysis technique, which gives the dimensions with the maximum variance. Using this information, we can filter out the dataset and operate on a reduced number of features. Post removal of columns, K-Means clustering algorithm is applied on the dataset to segment the customer data. To evaluate the appropriate value of K, we use the elbow plot.

### Phase 2: Customer identification with supervised algorithms

The motive here is to train the model with a supervised algorithm and predict which individual is more likely to be converted into a customer. We split the data into train and validation. The datasets are cleaned, and the model is then trained. We then evaluate the model based on the validation set. To evaluate the model with metrics we consider various metrics such as Accuracy, Precision, Recall and AUROC. However, using Accuracy in our case would be ineffective due to imbalance in the dataset. The split for the data is 98:2 %. Thus, even if the model predicts all points as the dataset with 98% records, the accuracy will still say 98%.



Thus, we use a metric known as AUROC which uses both True Positives and False Positives. Since, both these cases are important for our model, using AUROC as the evaluation metric is ideal for this case.

Area Under Receiver Operating Curve (AUROC) is a metric calculating as a ratio of True positives against False Positives. A good performing model will have a AUROC value close to 1. Thus, higher the AUROC value, more accurate the model is.

## 2. Data Profiling

### 2.1. Data Exploration and Pre-processing

Before proceeding with any of the data processing steps, it is necessary to ensure that the dataset has the same dimensions and attribute values as mentioned. This was verified using `shape()` and `head()` functions. The next step is to explore the data and resolve any issues with the data. All the data pre-processing steps performed have been listed below:

#### 2.1.1. Different datatype information in the same column

While loading the dataset into variables, the system throws an warning on columns "CAMEO\_DEUG\_2015" and "CAMEO\_INTL\_2015" indicating the columns contain values with conflicting datatypes. The attribute values dataset which contains the description for these attributes indicate that these columns hold numerical values. Thus, we replace any conflicting observations with NaN value.

- Values 'X' and 'XX' are replaced with NaN values in the dataset

#### 2.1.2. Addressing the issues with LP columns

The data contains values in the columns prefixed with LP, which contain information regarding an individual's family, financial status, and their life. These columns contain the value '0' as a default wherever the data is not recorded, thus we replace it with NaN.

- Columns 'LP\_LEBENSPhase' and 'LP\_LEBENSPhase\_GROB' contain multiple dimensional information. Thus, we split these columns into multiple columns.
- Columns 'LP\_FAMILIE\_FEIN' and 'LP\_STATUS\_FEIN' contain information which can derived from subsequent columns suffixed with GROB. Thus, we drop these columns.

#### 2.1.3. Resolving Unknown Values

Lot of columns in the dataset contain unknown values as described by the attribute excel sheet. The entries as specified are used to replace these values with NaN values. In total there are 232 dimensions in the dataset with unknown values.

#### 2.1.4. Mutual columns

The given dataset is used to understand the mutual columns across both general population and customer data. The attribute excel sheet provided is used to assist this process. In total, there are 272 columns which are common between the general population data and the customers data.

#### 2.1.5. Encoding Features

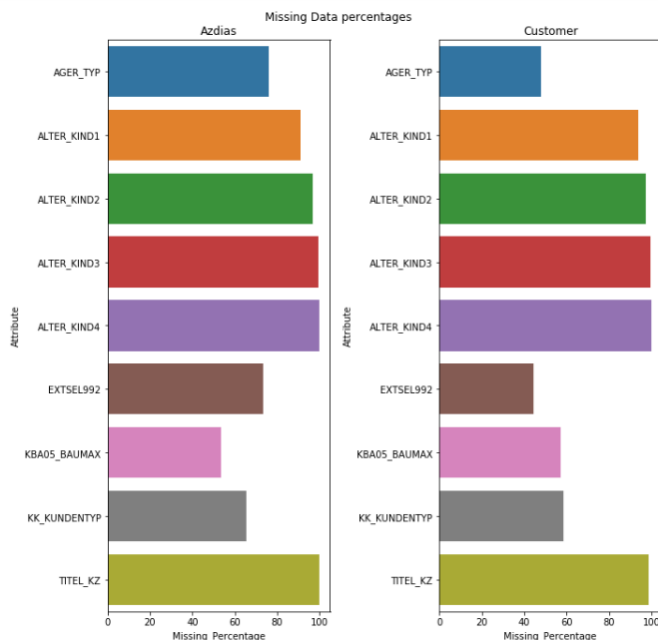
Since the dataset needs to be standardized before processing and training a model, it is necessary to encode the columns which do not have standard values.

- EINGEFUEHT\_AM: This column represents a date which could signify which date the record was created on. Since the data is represented as a string the source data, it is essential to convert this column to datetime.
- ANREDE\_KZ: This column represents the gender of the individual with values 1 indicating male and 2 indicating female. Since a normalized dataset contains values in the range 0,1 we change the values from 1,2 to 0,1.
- WOHNLAG: This column contains misrepresented values which are replaced with NULLs in our process
- CAMEO\_INTL\_2015: This column contains information about the status of a person. This column is divided into 2 columns to represent information about family status and wealth status separately.

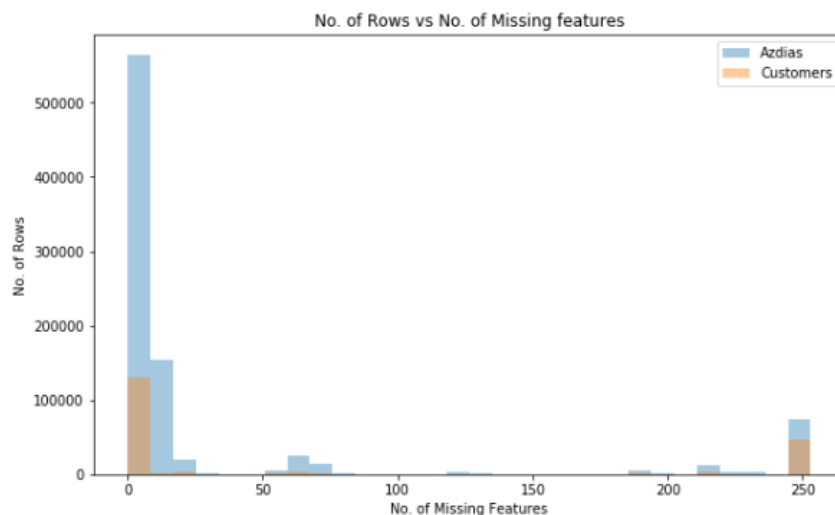
### 2.1.6. Missing Values

Finally, we deal with the missing values in the dataset. We look at the missing data in two perspectives, column wise and row-wise.

- Column-wise: The percentage of missing data is computed and plotted using a bar plot. Since more than 100 columns contain missing values, I decided on a threshold of 40% to restrict the data. Meaning, any column which contains more than 40% missing data will be discarded from further processes.



- Row-wise: Next, we look at the row values which contain missing values. We gauge the threshold by the number of features missing in each row. We set the threshold as 50 features and drop all the rows with more than 50 features missing. This resulted in a drop of 57406 in the customers dataset and a total of 153933 rows in the general population data.



### 2.1.7. Imputing Missing Values

An important step after removal of the columns and rows with a high number of missing information is to deal with the remaining rows and columns. Since the dataset represents customer and general population data, we replace the missing values with the most frequent values in the respective columns.

### 2.1.8. Feature Scaling

Finally, we need to scale the dataset between 0 and 1 before training the model. This is achieved using the standard scaler method in python.

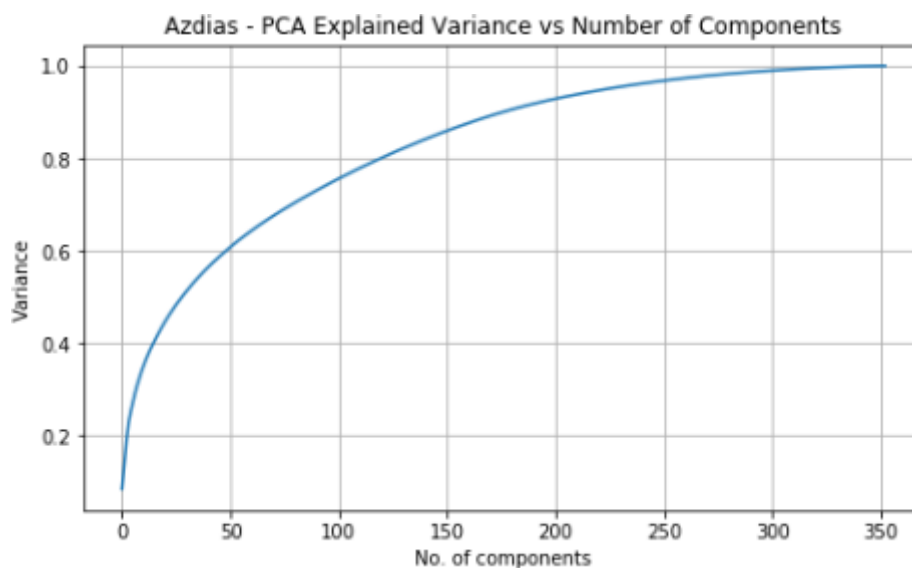
## 3. Algorithms and Techniques

### 3.1. Customer Segmentation

The goal here is to compare the customer and general population dataset and understand the patterns. Clustering the existing customers and the general population will help us understand which features are important for the customers. Comparing the customer dataset and the general population as it is would be a herculean task and much of the work will be redundant as not all attributes are important for training the model. Thus we use a technique known as Principal Component Analysis.

#### 3.1.1. Dimensionality Reduction

Using the entire data is not an ideal solution as not all attributes contribute to the model. Thus, a smart solution is to reduce the number of redundant dimensions and process only those features which contribute more. Principal Component Analysis carries out this task by ordering the features in the order of their significance. We thus, use this analysis to find out the number of features required to represent maximum variance in the dataset. We used this information to plot a graph as shown below:



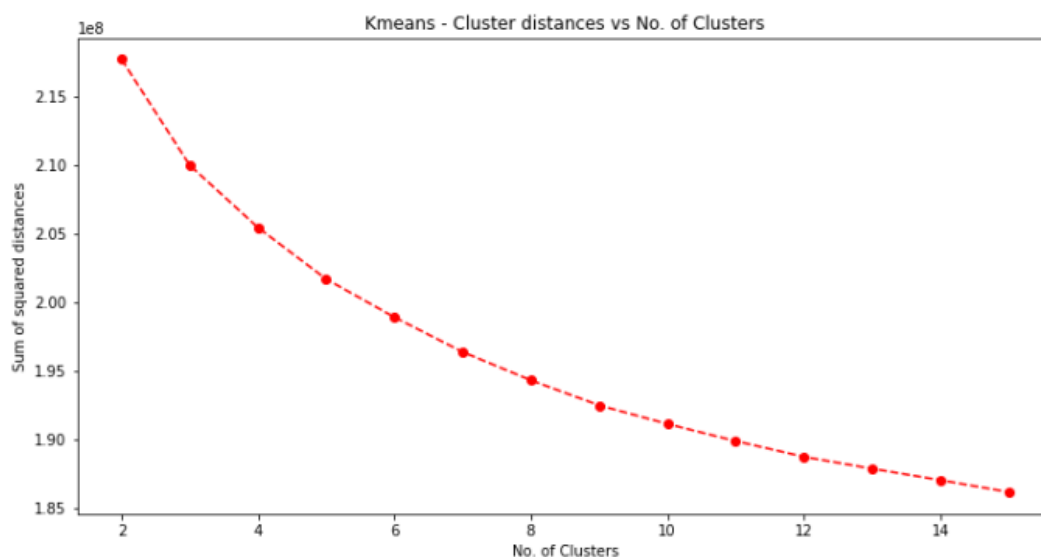
As seen in the figure above, amongst the 353 features originally present in the dataset, approximately 175 features are enough to explain 90% variance in the data. This means that, around 175 features cover 90% of the relevant information compared to the remaining features. We can therefore ignore the remaining features and consider these 175 features to proceed with our processing. Using the same Principal Component Analysis model, we filtered these features and created a new dataset.

#### 3.1.2. Clustering

Once the size of the dataset has been reduced with the dimensionality reduction technique, the next goal is to apply a unsupervised learning algorithm to segment the unlabelled data. This is achieved by using K-Means algorithm. It is a simple unsupervised learning algorithm used for clustering the data. It functions by assuming K cluster centroids and calculating the distance of each point from these centroids. It then assigns each point to the cluster where the distance between the point and the cluster centroid is minimum. After assigning each point to a cluster, it recomputes the centroids for the clusters by taking an average of all points belonging to that cluster.

Then, it recalculates the distance of each point to the cluster centroids and reassigns the points to the clusters. This is repeated until two consecutive iterations of cluster assignment are same. The goal in this model is to have maximum separation between different clusters while having a compact point assignment within clusters.

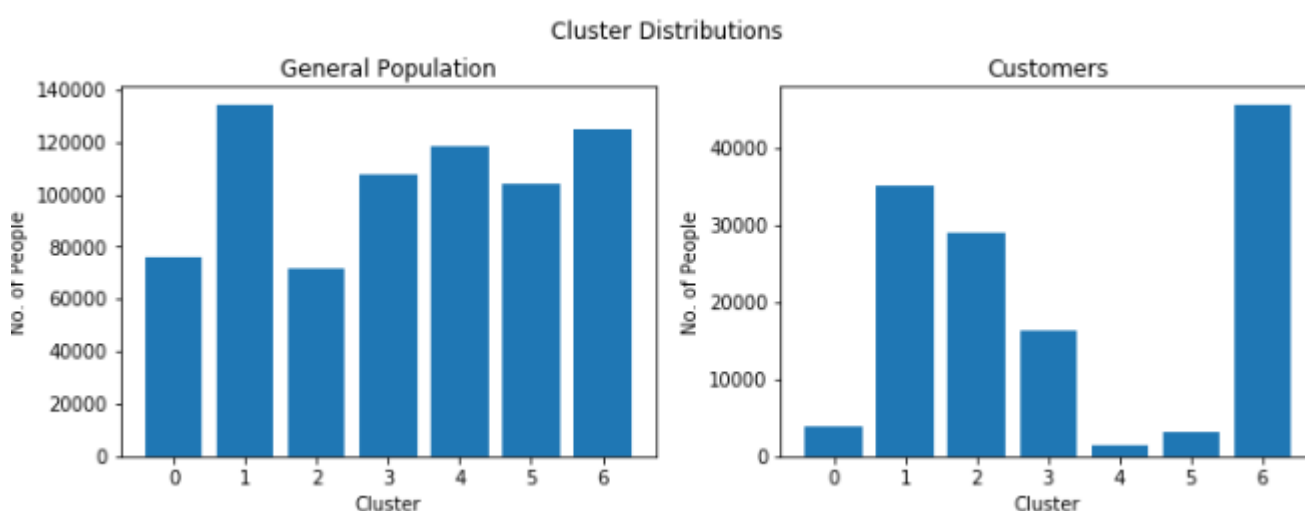
One challenge using this model is to select the value for K, as this is the key driving force for the entire calculations. An ideal value for K is neither too small nor too large. To decide the ideal value for our case, we use the elbow plot method as shown below:



K-Means Elbow plot

The idea of using this plot is to track the squared distances in each cluster for the respective value of K. As the value of K varies from 1 to 15, we see that the sum of squared distances decreases gradually. We select a value for K, post which, the sum of squared distances does not decrease by a high rate. Thus, as observed in the above graph, we use '7' as the value of K in our modelling.

Next, we formed 7 different clusters and analysed the distribution of general population vs customers. The figure below represents the number of individuals from these categories in each cluster.



As seen above, the general population is evenly distributed across 7 clusters, whereas, the customers are more in some clusters compared to others. Specifically, clusters 1,2,3 and 6 have a high number of customers compared to the remaining clusters 0,4,5. We can therefore infer that the individuals from general population in clusters 1,2,3 and 6 are more likely to convert into customers than the remaining clusters.



## 3.2. Customer Acquisition

After applying the data pre-processing techniques, feature extraction and an unsupervised learning model, the next goal is to use the test dataset provided by Udacity and applying a supervised learning model to predict the individuals likely to convert into customers. We use the file “Udacity\_MAILOUT\_052018\_TRAIN.csv” which contains similar features as the general population and customer datasets. Additionally, a column “RESPONSE” is provided, indicating whether the individual is a customer or not. We apply all the data pre-processing steps and feature extraction steps on the dataset provided before training our supervised model.

### 3.1.1. Benchmark Model

In order to gauge the performance of our final model, we need a benchmark model to compare our results to. This model is a basic model of logistic regression which is simple and efficient. The data is split into train and test and validated using this model. The AUROC score obtained through this model was 0.66. An AUROC curve is a plot of True Positive values and False Positive values where a value close to 1 is ideal. Thus, we will be using **0.66** as a benchmark for all our models.

### 3.1.2. Model Performance

After computing the benchmark results, we implemented different models to compare and select the best model for further processing. The source data was split into train and test datasets and used to gauge the performance of these models. We implemented the below models:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- Gradient Boosting Classifier
- AdaBoost Classifier

All these algorithms are classification algorithms which uses different decision boundaries to classify the data into different classes. Below results were observed upon implementing these models:

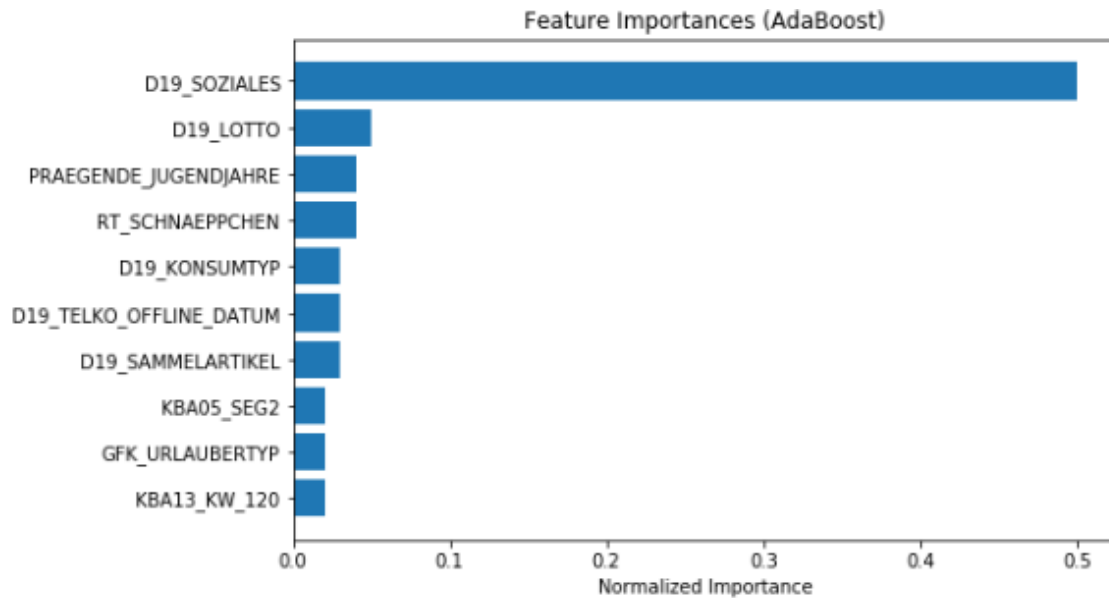
	Model	AUCROC_score	Time_in_sec
0	LogisticRegression	0.633664	15.9403
1	DecisionTreeClassifier	0.504843	3.85314
2	RandomForestClassifier	0.515099	1.55468
3	GradientBoostingClassifier	0.748781	45.7556
4	AdaBoostClassifier	0.715393	16.5517

As seen, both Gradient Boosting Classifier and AdaBoost Classifier have higher values compared to our benchmark model. Thus, we consider both these models. However, since the Gradient Boosting model operates on adjusting the gradient in each iteration, it is highly inefficient compared to the AdaBoost model. Therefore, we proceed with AdaBoost Classifier for further processing.

### 3.1.3. Hyperparameter Tuning

Grid Search method is used to improve the performance of AdaBoost Classifier. A range of values are selected for the model and the grid search implementation is used to select the best parameters for our model.

After implementing the model with the best parameter settings, we look at the feature significance :



This indicates that the feature “D19\_SOZIALES” is highly significant compared to the other models.

## 4. Conclusion and Results

We were able to implement a supervised learning algorithm to secure the objective of predicting which individual is more likely to be converted into customers. Through this project, we were able to

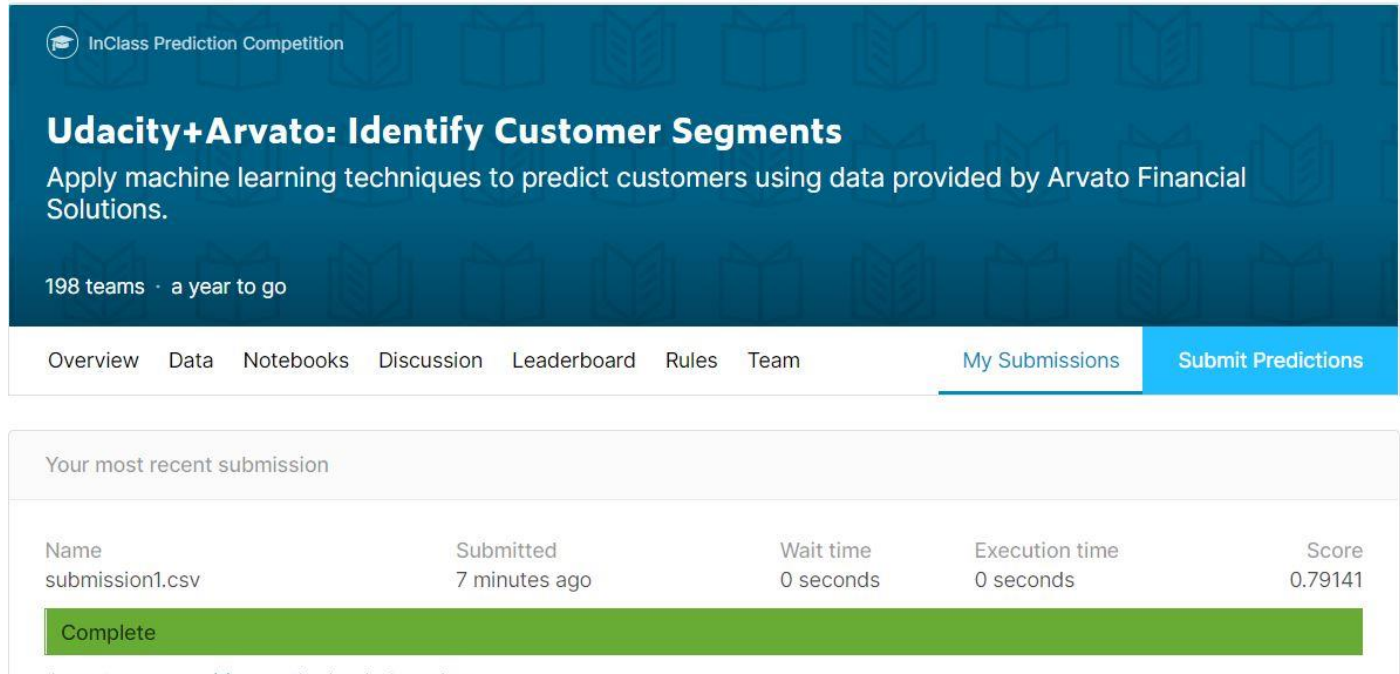
- Explore the data and implement data pre-processing steps
- Apply feature extraction technique known as Principal Component Analysis to reduce the dimensions in the source data
- Use an unsupervised algorithm to understand the distribution of customers in the general population and compare them
- Implement a benchmark model for the problem and evaluate it with an AUROC score – 0.66
- Finally, we were able to implement a supervised learning algorithm which is able to predict which individuals are likely to convert into customers, tune the model, as well as evaluate our final model

**Final Result: - AdaBoost Classifier= 0.7431**

## 5. Kaggle Submission

Final task as part of this project is to predict the results on the test dataset provided and submit the results to the Kaggle competition. The test file used is "Udacity\_MAILOUT\_052018\_TEST.csv". We performed all the data pre-processing steps, imputed and scaled the data. The best AdaBoost Classifier model was used to predict the results and create the submissions file.

The prediction file was submitted on Kaggle and results are shown below:



InClass Prediction Competition

### Udacity+Arvato: Identify Customer Segments

Apply machine learning techniques to predict customers using data provided by Arvato Financial Solutions.

198 teams · a year to go

Overview Data Notebooks Discussion Leaderboard Rules Team **My Submissions** Submit Predictions

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
submission1.csv	7 minutes ago	0 seconds	0 seconds	0.79141

Complete

[Link to your position on the leaderboard](#)

The resulting score on the date of submission was **0.79141**.

## 6. Future Scope

Although different data pre-processing steps were implemented in this project, there is still scope for improvement. Few of the steps which can be used to improve the results are listed below:

- Identifying more columns with categorical attribute information and encoding them
- Using a correlation matrix as a data pre-processing step to filter out features with low relevance
- Using data generation techniques to deal with the imbalance in the data

## References

- [1] Arvato-Bertelsmann, "Arvato", Bertelsmann, "<https://www.bertelsmann.com/divisions/arvato/#st-1>
- [2] Aditya Mishra, "Metrics to Evaluate your Machine Learning Algorithm", TowardsDataScience, 2018, "<https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234>"
- [3] AlindGupta, "Elbow Method for optimal value of k in KMeans", GeeksforGeeks, 2020, "<https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/>"