Capstone Report

Customer Segmentation and Optimization of Customer Acquisition with Arvato Financial Solutions

Kibru Temesgen 4/4/2023

AWSMachine Learning Engineer Nanodegree School of Artificial Intelligence

CONTENTS

1. DEFINITION

Project Overview

One of Udacity's suggested projects for the Machine Learning Engineer Nanodegree Program was "A Customer Segmentation Report for Arvato Financial Solutions." The project's primary goal is to identify the descriptive profile of the possible client and the likelihood that a new individual from

A new consumer can sign up as a result of the targeted mailout effort.
Arvato has made available a number of dataset files with data on the demographics of the general German population, existing Arvato clients, the results of targeted mailout campaigns, and two files with a description of the demographic traits.

The project is contains the following parts:
1. Data Analysis and Preprocessing;
2. Customer Segmentation Report;
3. Supervised Learning Predictive Model;

Domain Background

Arvato is a wholly-owned subsidiary of Bertelsmann. It is a global services company that provides a wide range of services such as customer relationship management, supply chain management, financial services, and IT services1. Arvato has over 70,000 employees in more than 40 countries worldwide.

Arvato provides mail services such as direct mail marketing, transactional mail, and hybrid mail. They also provide e-commerce fulfillment services such as warehousing, order management, and shipping. Arvato's mail services are designed to help businesses manage their customer communications more efficiently and effectively.

In order to help its clients make business choices, Arvato provides them with useful data insights. One of the industries that is expanding is customer-centric marketing. analyzing the data to find hidden trends and client behavior. Arvato is offering insightful information to businesses engaged in customer-centric marketing. Nowadays, data science and machine learning are widely employed to achieve company objectives and please consumers.

In this project, Arvato is working with a mail-order business that offers organic items in Germany to better understand its consumer demographics and find potential new customers. To better comprehend the various client categories, demographic information about the German population will be analyzed, and a system will then be built to forecast whether or not an individual would become a customer based on that information.

Datasets and Inputs

Four datasets are provided for completion of the project along with additional two metadatas about the attributes of the 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).

2 metadata files have been provided to give attribute information:

- 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

In the framework of the Machine Learning Nanodegree Program, Arvato has made all of the project-related information available for analysis and client segmentation. The four csv files are the demographic data files, and each row in them contains information about a specific person's demographics. In addition to their demographics, each row also contains extra details about their home, building, and neighborhood. Three extra columns in the customer data list their details in relation to the mail order business. For the purpose of evaluating supervised learning algorithms, Train and Test data have been made available.

Problem Statement

The problem can be stated as "How can a mail order corporation efficiently recruit new consumers given the demographic data of a person?"

With the use of unsupervised learning algorithms, the demographic information of both the general population and the clients is first investigated. Finding out the demographic characteristics associated with someone being a client of the mail-order firm is the aim of this stage, which also involves identifying segments in the general population and segments in the existing customers.

On the basis of the demographic information, a supervised learning algorithm is then employed to estimate whether a person is likely to be a client or not.

Evaluation Metrics

AUC ROC is a good metric for classification problems with imbalanced data. The ROC AUC is sensitive to class imbalance in the sense that when there is a minority class, you typically define this as the positive class and it will have a strong impact on the AUC value. This is very much desirable behavior. The True Positive Rate against False Positive Rate is shown using the AUC-ROC. AUC of 1 indicates that True Positives and True Negatives are completely distinct from one another, whereas AUC of 0 indicates that the models classify all True Negatives as True Positives and the other way around.

Area Under Receiver Operating Characteristic (AUROC) has been chosen as an evaluation metrics because of this. The True Positive Rate and False Positive Rate are plotted under various threshold settings to construct the AUROC curve, which provides information about the overall performance of the model.

2. ANALYSIS

DATA EXPLORATION AND PRE-PROCESSING

I. Data exploration

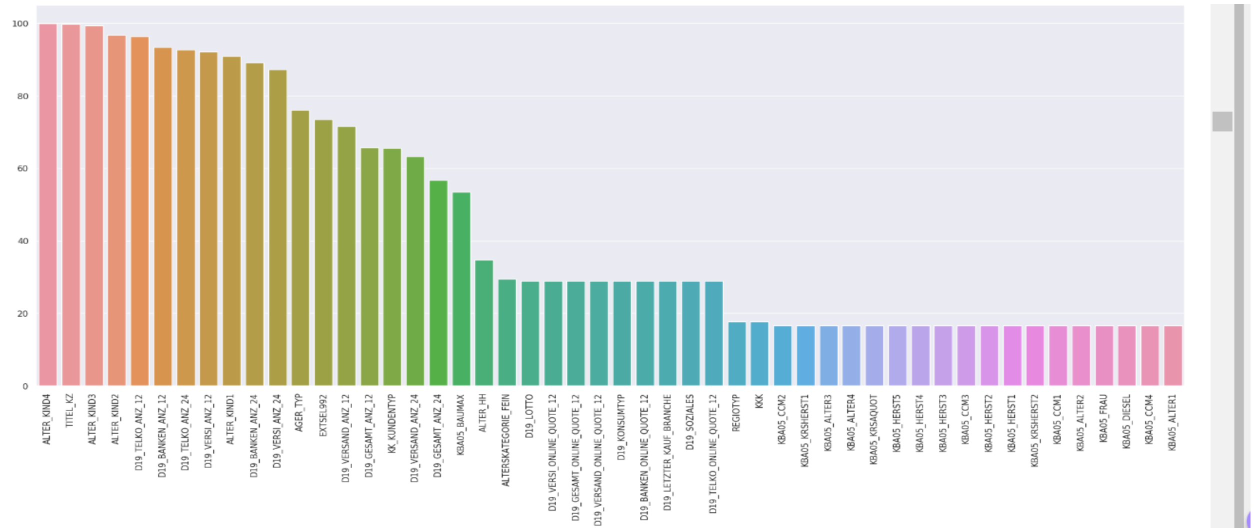
The population and customers datasets are loaded and checked to contain the information expected. The number of rows and columns are also checked along with descriptive statistics such as mean, std, freq, and etc.

II.Converting Uknown values per each columns into Nan

The first step is to address unknown values. Here, uknown values exist as Nan values and additionally each column has their own designation for unknown records. Hence, the first task is to figure out, for each column, what is used to represent unknown values and change those values to Nan. The Dias Attribute file containes this information. This file is used to find the representation of unknown values(other than Nan) for each column. After finding this values, each of the records containing these values are replaced by Nan.

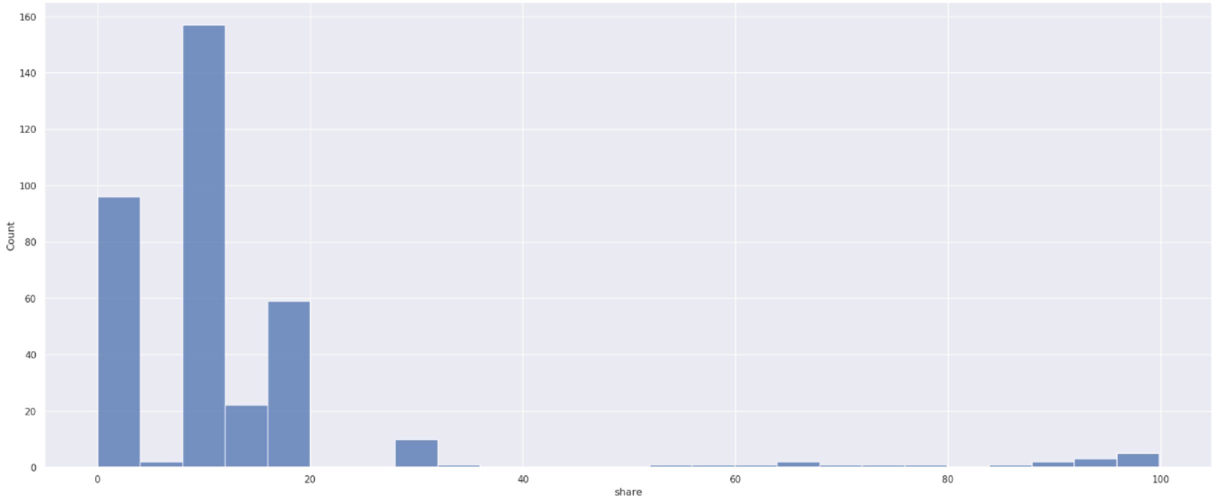
III. Dropping columns after analysing percentage of Nan per columns

Here, the composition of the dataset will be explored and visualized. The percentage of Nan values will be analysed for each row and column to decide between imputation and dropping of the column/row.



we can observe that among 365 columns, more than 20 columns has 20% or more of their records missing. For the rest of the columns missing values are less than 20%.

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f88719341d0>
```

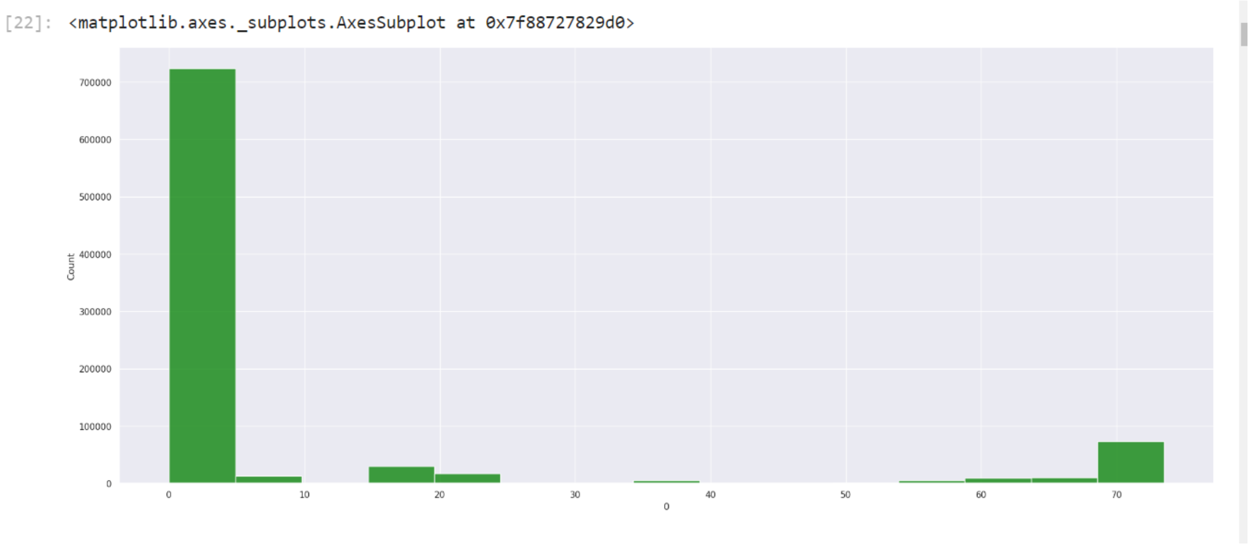


If we visualize the percentage of unkown values with respect to the number of columns. we can see that the majority of columns has only less than 40% of their records missing.

Hence, it is decided to drop columns that has more than 40% of their records missing since it doesnt result in significant loss of the information contained in the dataset

III. Dropping rows after analysing percentage of Nan per columns

The majority of the rows has only less than 10% of their columns missing hence, we can use 10 as a threshhold for dropping.



```
[22]: <matplotlib.axes._subplots.AxesSubplot at 0x7f88727829d0>
```

IV. Analysing and manipulating categorical columns

The following columns contains categorical variables. 'D19_LETZTER_KAUF_BRANCHE', 'CAMEO_DEU_2015', 'CAMEO_INTL_2015', 'CAMEO_DEUG_2015', 'EINGEFUEGT_AM', and 'OST_WEST_KZ'

- CAMEO_DEU_2015 contains detailed information about the population. The information in CAMEO_INTL_2015 is contained in CAMEO_DEU_2015 hence removing this column won't result in significant loss of important information.Hence, this column would be dropped.
- The D19_LETZTER_KAUF_BRANCHE attribute containes 35 categories. Since the essence of this attribute is not given in the DIAS file, it is better to drop it. We dont know what the categories represent and there is too many of them which complicates the dataset.
- OST_WEST_KZ is simple as it containes two categories and can be encoded into zero and one.

- EINGEFUEGT_AM is a datetime attribute encoded as categorical.Hence, we would convert it to datetime column.
- CAMEO_DEUG_2015 containes 18 categories which can be converted to integers.
- Replacing X and XX values with NAN in CAMEO_DEU_2015 and CAMEO_DEUG_2015

The above process is repeated for the customers dataset

V. Feature Engineering, Imputing Nan values, One hot encoding, and Scaling of features

In this section nan values will be imputed with proper values and categorical values would be one hot encoded. Columns with numerical values would be scaled to allow for fast convergence of the later algorithms.

- ☐ Nan values in categorical column will be imputed by most frequent values.
- ☐ Nan values in numerical columns will be imputed by the median
- ☐ Nan values in binary column will be imputed by most frequent values.
- ☐ Categorical columns will be one hot encoded
- ☐ Numerical columns will be scaled using standard scaler

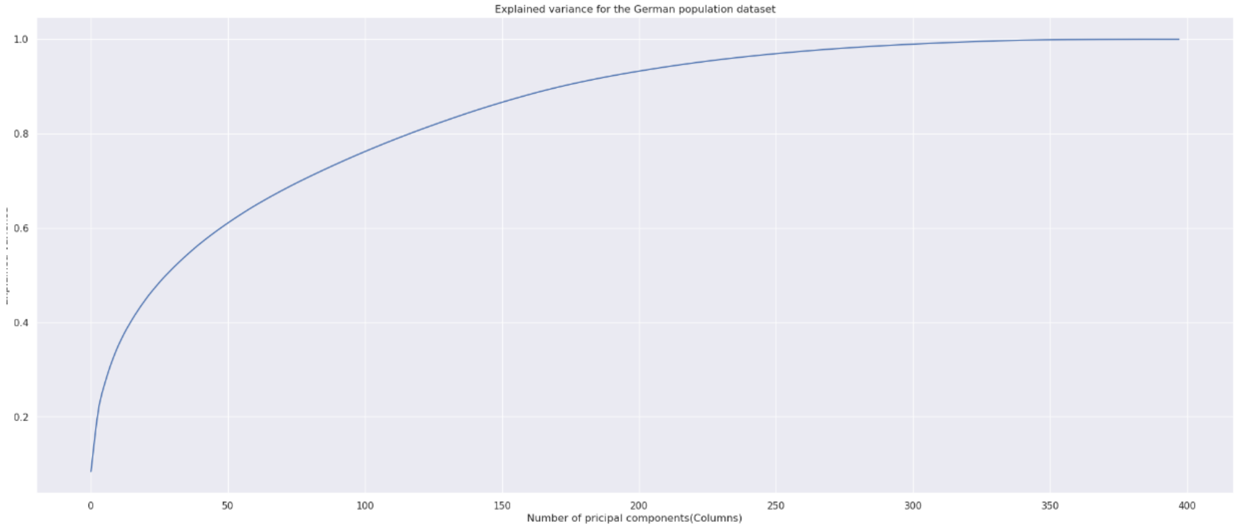
The same procedure is repeated for the customers dataset

VI. **Dimensionality reduction and principal component analysis(PCA)**

AS we have more than 390 features, it is important that we reduce the number of features by retaining only the essential parts. PCA is a widely used technique to do this task.

PCA can help by reducing the number of features in the dataset while retaining the most important information. It does this by finding the directions of maximum variance in the data and projecting the data onto a lower-dimensional space. This new space has fewer dimensions, which makes it easier to analyze the data and find patterns or relationships.

By analyzing the principal component analysis of our data set we can decide how many features to keep.



From the above plot, we can see that more than 90% of the variance can be explained with just 200 features. Hence, we can take the 200 most essential attributes while retaining more than 90% of the information. The trade off is beneficial as we could reduce the dataset in half while only losing less than 10% of information.

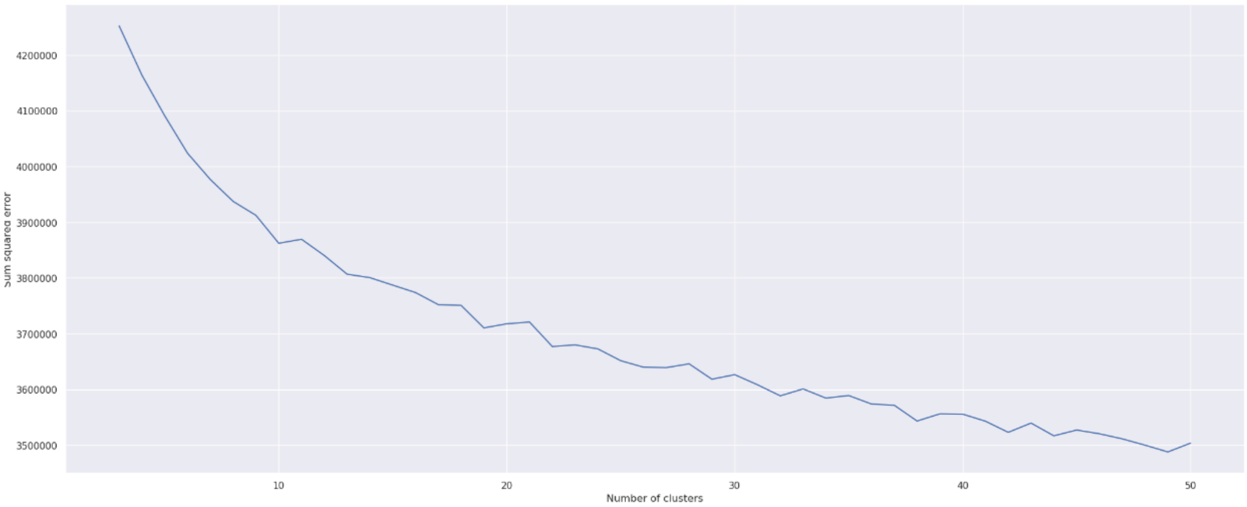
VI . Algorithms, Techniques, and Methodology

A. Customer Segmentation

Characterizing the link between current consumers and the demographics of the German population is the primary objective of the project's first phase. It also aims to identify the general population's demographic characteristics that are most likely to lead to customers.

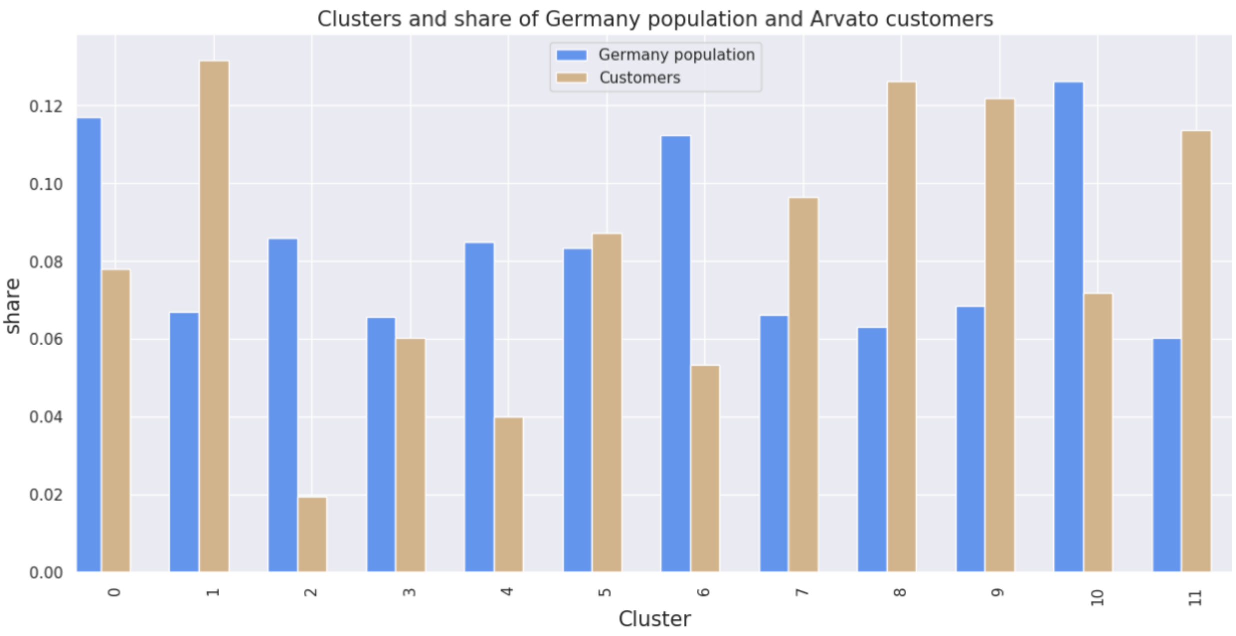
1. clustering

The next step is to segregate the general population and the client base after PCA characteristics have been set up. For the purpose of segmenting the general public and customers, the unsupervised clustering method k-means is used. In order to properly segregate the data, we want to choose a k such that data points in a single cluster are near to one another while yet having a sufficient number of clusters. The centroid distance often reaches a "elbow" after experimenting with several values for k; at this point, the pace of its decline abruptly ceases, indicating a suitable value for k.



As the dataset is complex and contains many features, it is not completely clear where the elbow exists. However, it can be observed that there was a steep decrease in the squared error until around cluster number 12. Hence, 12 will be taken as the optimum number of clusters.

We next compared the general populace and customer base for each of the 12 clusters as we turned our attention to clusters. We can determine which clusters had excessive consumer representation and which ones didn't. This helps us think about how to better target potential consumers.



Relatively, the customers dataset has more variation among clusters than the german dataset. Some clusters of the population are over represented while other are under represented.

- cluster 1, 9, 11, and 8 are highly overrepresented in the customers while cluster 7 is not that much overrepresented. Hence, this clusters should be targete on the campaign as they are highly likely to be our new customers.
- cluster number 0, 2, 4, 6 and 10 are underrepresented in our customers. Hence, future campaing should avoid targeting this clusters as this segment of the population is not interested our product.

B.Supervised machine learning

In the second half of this research, potential clients will be predicted using supervised learning techniques based on demographic information. The MAILOUT train and test datasets will be used. As previously discussed, the MAILOUT dataset had a further column added to the training dataset that represented a person's likelihood of becoming a client of the business. Similar pre-processing and scaling techniques used for AZDIAS and CUSTOMERS were also used for the two datasets.

With the MAILOUT TRAIN dataset, we also saw significant class imbalance . We resampled the minority class to address this, resulting in an equal number of observations in both answer groups.

1) Benchmark Model: Logistic Regression

Setting a benchmark, which represents the baseline performance using a straightforward model, is the first stage in supervised learning. The performance of other trained models will be compared to and evaluated against the benchmark model. A common benchmark model is the logistic regression model.

After training the logistic regression model, an accuracy of 82% was obtained. This result would be used as a benchmark.

2, AutoGluon

AWS AutoGluon is a powerful tool that can be used to train machine learning models for tabular data prediction and more. It can automatically handle many of the tasks that are typically associated with model training, such as feature selection, hyperparameter tuning, and model selection.

AWS AutoGluon can try multiple models and pick the best one. It does this by automatically selecting the best algorithm for the data and then tuning the hyperparameters of the algorithm to maximize performance. AutoGluon can also try multiple combinations of algorithms and hyperparameters to find the best performing model.

For these reasons, AutoGluon would be used to train different models and pick the best one with acceptable evaluation metrics value.

The following model was trained using Autogluon.

- KNNModel: [3]A supervised machine learning algorithm that classifies data points by finding the nearest neighbors in the training data. It is a simple and effective algorithm that is often used for classification tasks.
- LGBModel: [4]A supervised machine learning algorithm that uses gradient boosting to classify or regress data. It is a powerful algorithm that can achieve state-of-the-art results on many tasks.
- WeightedEnsemble_L2: [5]A supervised machine learning algorithm that combines the predictions of multiple models using a weighted average, where

the weights are determined by the L2 norm of the residuals. It is a robust algorithm that can improve the performance of individual models.

- CatBoost is an open-source machine learning library that uses gradient boosting on decision trees. It is known for its speed, accuracy, and ability to handle categorical features.

Model Type	**score**
WeightedEnsemble_L2	0.999861
CatBoost	0.999804
LightGBMXT	0.999768
LightGBM	0.999680
KNeighborsDist	0.990800
KNeighborsUnif	0.990800

Over all, all the models have scored a very good roc_auc value. The best model can be taken as WeightedEnsemble_L2 .

3. JUSTIFICATION AND RESULT PREDICTION

The selected models perform better than the benchmark model which achieved a score of 82%, but these models were able to achieve scores of above 99% which makes it acceptable.

By picking the best model, a prediction has been made on the test data set. 221 potential new customers have been identified. This customers can be targeted for the campaign as they are likely to respond.

4. FUTURE RECOMMENDATIONS

In the future, there are room for improvements and better accuracy of the data set is analysed in a much deeper detail and different algorithm can be tried.
Since the dataset contains more than 300 features, much can be achieved if each features are considered thoroughly and their effect is analysed. Rather than simply

applying a PCA algorithm to reduce features, a more informed and indepth study of the variables would result in a much better result and accuracy.

More algorithms can be tried to get a better accuracy.

REFERENCES

[1] Arvato. Wikipedia. https://en.wikipedia.org/wiki/Arvato
[2] Customer Segmentation. Wikipedia.
https://en.wikipedia.org/wiki/Market_segmentation

[3]KNNModels
https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
[4]LGBModel https://lightgbm.readthedocs.io/en/latest/Parameters.html
[5] WeightedEnsemble https://pypi.org/project/WeightedEnsemble/