Arvato-Bertelsmann Customer Segmentation Report

Udacity – Machine Learning Engineer, Nanodegree
Capstone Project

Aakriti Sharma

October 30th, 2020

ABSTRACT

Maintaining healthy customer-relationships is of utmost importance for any company today. In this project, we predict customers for a mail-order sales company based on demographics of general population and the company's customers. Using unsupervised learning techniques, we perform customer segmentation to help us identify the potential buyers, target the most profitable customers, and retain them over time. Then we apply these learnings and predict the individuals who are more likely to respond to the company's sales campaign.

1. DEFINITION

1.1. Project Overview

This project was made towards completion of Udacity's Machine Learning Engineer Nanodegree. The goal of this project is to help a mail-order sales company in Germany to identify segments of the population to target with their marketing campaign to acquire new customers.

To help with this problem, Arvato has provided us data on general population demographics, their customers and their client response to previous campaign. There are four data files associated with this project:

- **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).

Using these datasets, we have to predict which segments of population to target, in order to target the right audience.
The solution to the above-defined problem is divided into two parts. First, comparing the general population with the existing customers and identifying the segments of the general population that are more likely to be part of the company's main customer base, with the help of unsupervised learning techniques. Second is building a supervised learning model to predict the response of individual customers towards the marketing campaign.

1.2. Problem Statement

The goal of this project is to help a mail-order sales company in Germany to identify individual and segments of the general population to target with their marketing campaign. The company has provided us with the demographic data of their current customers and the general population. We've to build a customer-segmentation and acquisition machine learning model for the company, which correctly categorizes customers into groups and identifies the customers that the company should target.

1.3. Evaluation Metrics

The evaluation metrics used for measuring the performance of our classification model is Area Under the Curve Receiver Operating Characteristics. This metric is helpful for multiclass classification, along with that, it takes care of class imbalance. Since most individuals don't respond to the mail out, it is obvious that the number of positive responders will be way less than those who responded negatively or didn't responded.

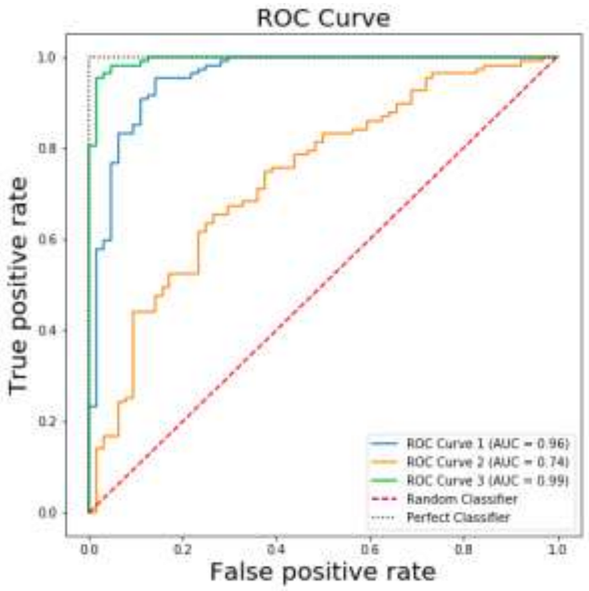


Figure 1: Graphical Representation of AUROC Curve

2. ANALYSIS

2.1. Data Exploration and Exploratory Visualizations

The demographics data files consisted of 366 features containing information about outside of individuals, their household, building and neighborhood. The customers file also consisted of three additional features ‘CUSTOMER_GROUP’, ‘ONLINE_PURCHASE’, and ‘PRODUCT_GROUP’, which provided broad information about customers and the mailout-train file had 1 additional feature ‘RESPONSE’, which indicated whether or not each recipient became customer of the company.

	Row_1	Row_2	Row_3	Row_4	Row_5
LNR	910215	910220	910225	910226	910241
AGER_TYP	-1	-1	-1	2	-1
AKT_DAT_KL	NaN	9.0	9.0	1.0	1.0
ALTER_HH	NaN	0.0	17.0	13.0	20.0
ALTER_KIND1	NaN	NaN	NaN	NaN	NaN
ALTER_KIND2	NaN	NaN	NaN	NaN	NaN
ALTER_KIND3	NaN	NaN	NaN	NaN	NaN
ALTER_KIND4	NaN	NaN	NaN	NaN	NaN

Table 1: sample from Udacity_AZDIAS_052018 dataset

Missing Values: Some values in the data files represented unknown or missing data points. After identifying and replacing those missing data values from our data sets, it was found that many columns had more than 30% of the data missing in both customers and general population files. It was also identified that 11.22% (99968) rows in general population and 26.50% (50786) in customers file had more than 50% of the feature values missing.

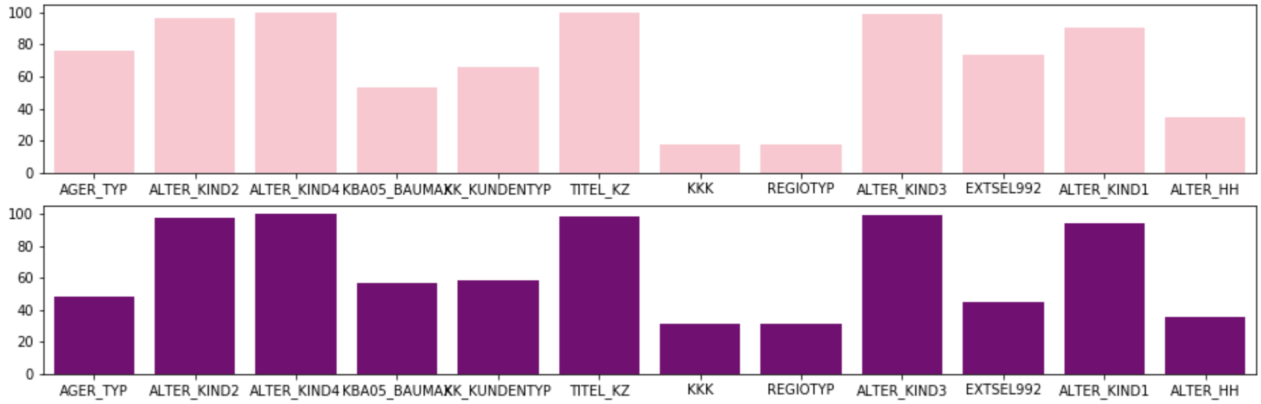


Figure 2: Comparison of missing data in General population (graph 1) and customers (graph 2) column-wise

Customers: On analyzing customers data, we identified few characteristics of the customers. Only 9.02% of the total customers made online purchase and 31% of all the customers were single buyer.

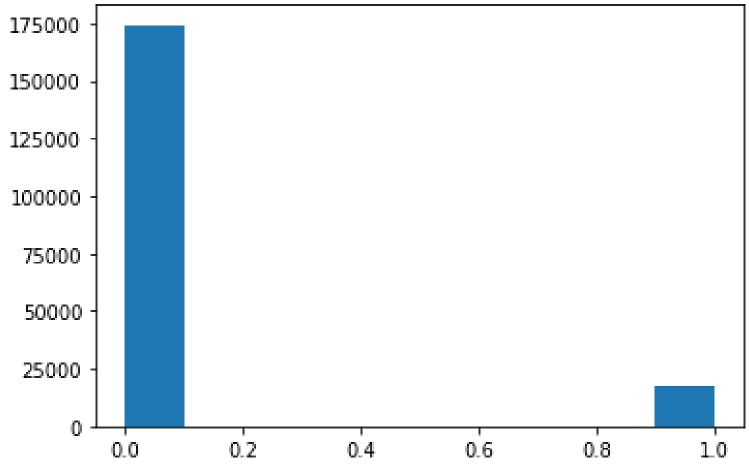
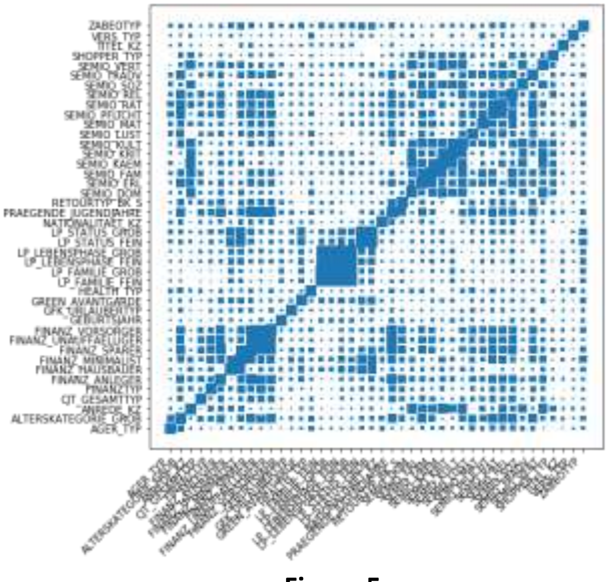


Figure 3: Number of customers making offline purchase vs online purchase

Information Levels and correlation: The features were broadly divided into 10 information levels (PLZ8, Microcell (RR3_ID), Person, 125m x 125m Grid, Household, Building, Community, RR1_ID, Postcode, Microcell (RR4_ID)). On plotting features according to these levels, the plot showed large-boxy patterns representing that group of various features in dataset were highly correlated to each other.



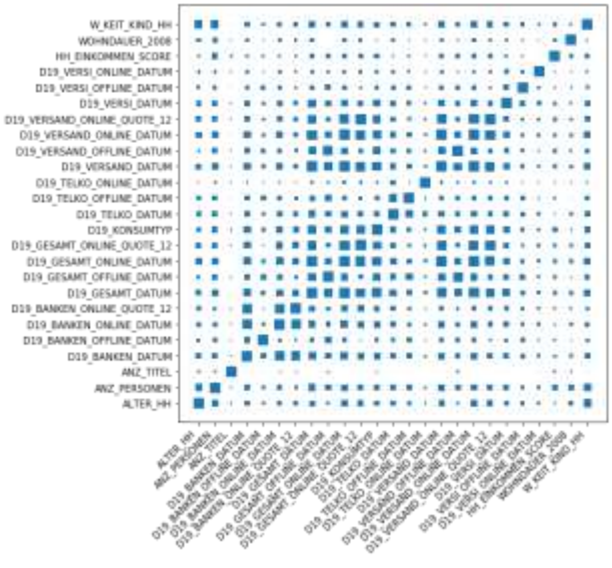


Figure 5 **Figure 5.2**

Figure 5.1: Correlation heatmap of 'Person' level attributes.

Figure 5.2: Correlation heatmap of 'Household' level attributes.

2.2. Feature Encoding and Engineering

Initially, columns 'CAMEO_INTL_2015' and 'CAMEO_DEUG_2015' gave ‘Mixed datatype’ warning due to mixture of integer type of data and ‘X’ and ‘XX’ strings representing missing data. After the replacement of values representing missing data, these columns were converted back to integer type.

Columns ‘LP_STATUS_GROB’ and ‘LP_FAMILIE_GROB’ had various values representing same category, so these columns were re-encoded, such that each category was represented by a single unique value. After this, we extracted year from ‘EINGEFUEGT_AM’ datetime column.

Column ‘CAMEO_DEU_2015’ represented the similar data to ‘CAMEO_DEUG_2015’, so the former feature was removed from the dataset, along with columns having more than 30% missing values.

Categorical features left after all the conversion were one hot encoded and after simple imputer was used to impute missing data with most frequent value in the column. At last the all the columns, except ‘LNR’, were scaled using Standard scaler technique.

2.3. Algorithms and Techniques

2.3.1. Unsupervised Modelling

In this first part of our project, our goal was to identify groups of people based on their demographic information. Here, we made use of unsupervised learning techniques to describe the population that is most likely to be part of mail-order company’s primary customer base.

Principal Component Analysis: PCA is a fundamentally a simple dimensionality reduction technique that transforms the columns of a dataset into a new set features called Principal Components (PCs). The information contained in a column is the amount of variance it contains. The primary objective of Principal Components is to represent the information in the dataset with minimum columns possible.

K-means Clustering: k-means clustering tries to group similar kinds of items in form of clusters. It finds the similarity between the items and groups them into the clusters. K-means clustering algorithm works in three steps:

1. Select the k values.
2. Initialize the centroids.
3. Select the group and find the average.

2.3.2. Supervised Modelling

After we’ve found which parts of the population are more likely to be customers of the mail-order company, we move to build a prediction model. With the help of demographic

information of each individual, we’ve to decide whether or not it will be worth it to include that person in the campaign.
To perform this task, we used various classification models and finally build a Light XGBoost classifier.
LightGBM, short for Light Gradient Boosted Machine, is decision tree-based framework. It supports various algorithms. It has XGBoost’s advantages, along with sparse optimization, parallel training, multiple loss function, bagging, regularization, etc.

2.4. Benchmark Model

We used AUC metric to measure the performance of our classifier. The model with highest AUC on cross-validation set was considered to be the best model. We trained Logistic regression model, with default parameters, to predict the response for the test and train sets. The obtained AUC score was considered as base score.

3. Methodology

3.1. Data Preprocessing

3.1.1. Parsing Data

The first step of preprocessing was to parse the data with correct datatypes and to identify the missing data values. To do this we used the meta data provided by Arvato and made the necessary replacements. Then, we dropped the columns which had more than 30% missing values in both general population dataset and customers dataset.

3.1.2. Column Transformations

After parsing the data, we converted our one hot encoded our categorical data and then Imputed the missing values with most frequent values using sci-kit learn’ s Simple Imputer.
At last we scaled our data using Standard Scaler, for the steps to be performed further.

3.2. Implementation

3.2.1. Principal Component Analysis

Before clustering our population, we performed dimensionality reduction of our dataset using PCA (Principal Component Analysis) and reduced the number of columns to 180 components while preserving 87% variance of original data. We chose the number of components based on the cumulative explained variance of components.

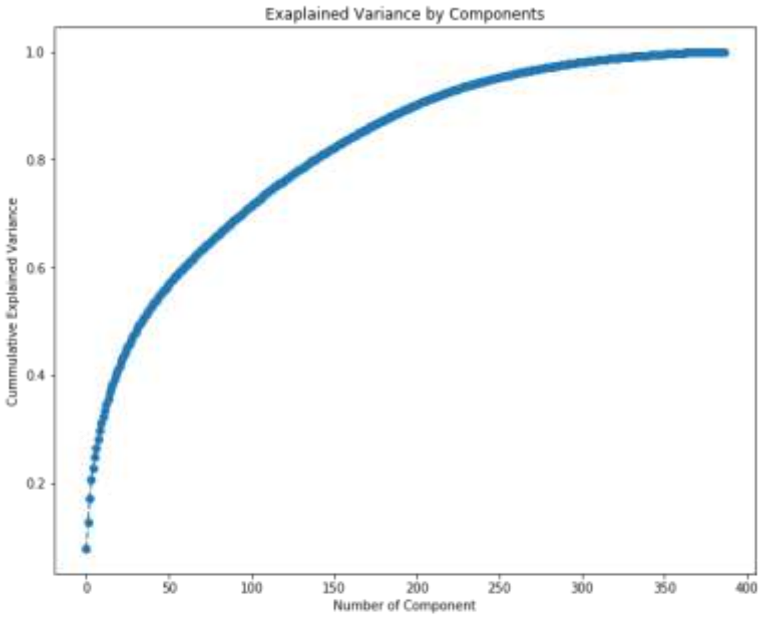


Figure 6: PCA explained variance by number of components.

3.2.2. K-Means Clustering

The first step of building a K-means clustering model is to define the value of K. To do this, we trained K-means model on sample of 20,000 rows of our reduced dataset for different values of K ranging from 1 to 20. To select optimal value for K, I performed Elbow method, by plotting values of K against WCSS (within cluster squared sum).

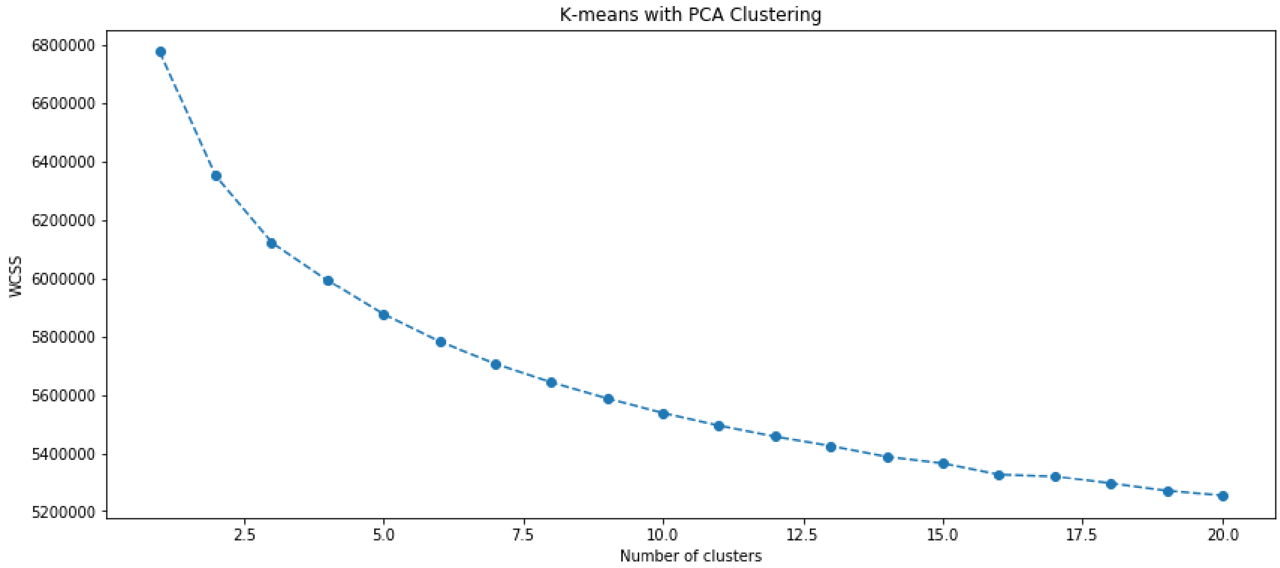


Figure 7: Elbow Plot

After choosing the value of K to be 5 from the above plot, we trained the k-means model on complete dataset. On clustering our population, it was found that certain segments contribute less to our customer-base. We can clearly notice that more than 50% of our customer population belong to cluster 0 and 2.

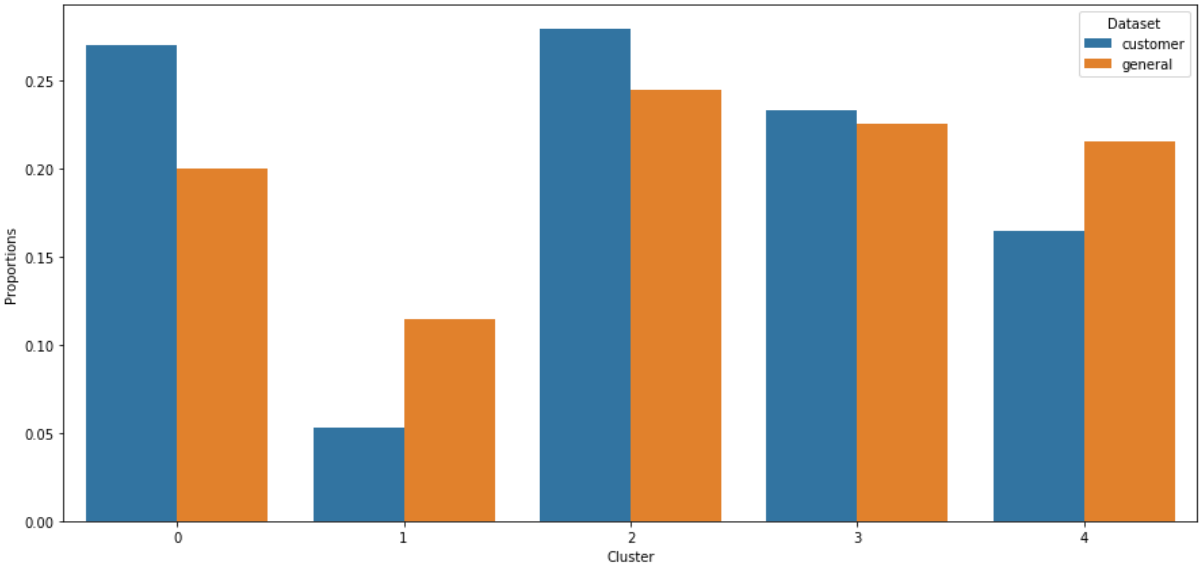


Figure 8: Distribution of customers and general population among the five clusters

3.2.3. Supervised Learning

To implement supervised learning model, we first processed our mailout train and test dataset through our previously defined pipeline. Then we performed dimensionality reduction and clustering to get the groups for the mailout dataset population.

On comparing the positive and negative response from train dataset, it was found that there is a class imbalance, with only 1.24% of people responding to campaign.

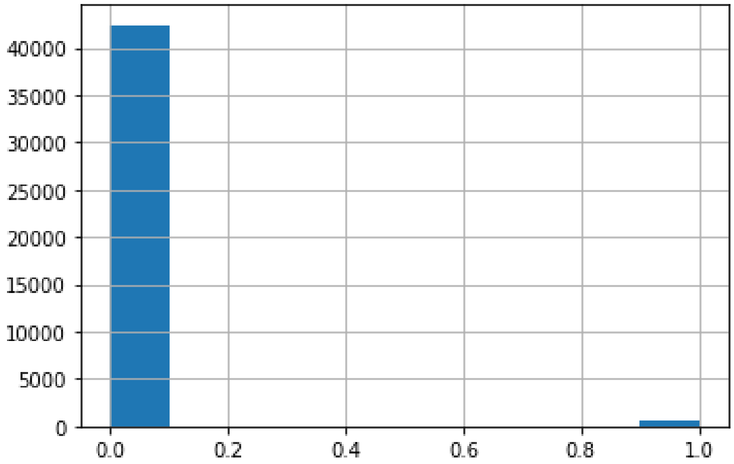


Figure 9: Distribution of customer responses

To deal with imbalance in data we used AUC metric to evaluate the model. We split the mailout train dataset using Stratified K-fold technique (where k was equal to 5.) We trained a few classification models to select one of them and refine it later.

	Train Score	**Validation Score**	**Std. Deviation**
Logistic Regression	0.79	0.65	0.010
Random Forest	0.99	0.60	0.535
Gradient Boost	0.92	0.74	0.015
XGB Classifier	0.88	0.74	0.032
LGBM Classifier	0.89	0.71	0.012

Table 2: AUC Score for various models

Based on these results, I chose to use LGBM Classifier as it’s faster to train and can be tuned easily over multiple iterations.

3.3. Refinement

To tune our Classifier, we used Bayesian parameter search over a hyper-space for the following parameters:

- 'learning_rate': (0.01, 0.1, 'uniform')
- 'num_leaves': (2, 180)
- 'max_depth': (2, 8)
- 'colsample_bytree':(0.5, 1.0, 'uniform')
- "min_data_in_leaf":(20, 100)
- 'min_child_samples': (0, 40)
- 'max_bin': (100, 600)
- 'reg_lambda': (1e-9, 1.0, 'log-uniform')
- 'reg_alpha': (1e-9, 1.0, 'log-uniform')
- 'scale_pos_weight': (10,90, 'uniform')
- 'n_estimators': (10, 200)
- ‘application’: 'binary'
- ‘Metric’:'auc'
- ‘n_jobs’:-1

4. Results

4.1. Model Evaluation and Validation

The final model performed well not only on training and validation set but also on the test set. The final model scored 0.79634 on test data on Kaggle, which is way better than random guessing (0.5 score).

5. Conclusion

5.1. Reflection

The success of any mailout marketing campaign depends on the customer list to which it is being targeted (Bult and Wansbeek, 1995; Bult et al., 1997; Dan Van Poel, 2003). Hence, it is crucial to strategically group customers based on various factors and target the correct audience.

Through this project we identified the population which might respond positively to the marketing campaign. We analyzed the demographics of the general population and the customers, and identified the features which best describe the customer-base.

5.2. Improvements

The current model placed me at 106th rank on Kaggle, which certainly means there's a room for improvement. In future, I plan to tweak and manipulate certain areas of the project, like:

- Engineering categorical features
- Re-defining missing or unknown data values.
- Using some supervised learning model to predict the missing data.
- Applying other techniques to balance data

References

[1] Dirk Van den Poel, 2003, Predicting Mail-Order Repeat Buying: Which Variables Matter? Available at: http://ebc.ie.nthu.edu.tw/km/MI/crm/papper/wp_03_191.pdf

[2] Selva Prabhakaran, Principal Component Analysis[PCA] – better explained: https://www.machinelearningplus.com/machine-learning/principal-components-analysis-pca-better-explained/

[3] Analytics Vidya: https://www.analyticsvidhya.com/blog/2020/10/a-simple-explanation-of-k-means-clustering/

[4] LGBM Wikipedia: https://en.wikipedia.org/wiki/LightGBM

[5] Scikit-Learn Docs. Available at: https://scikit-learn.org/stable/modules/clustering.html#silhouette-coefficient

[6] Figure 1: https://vitalflux.com/roc-curve-auc-python-false-positive-true-positive-rate/