Domain background:

Customer segmentation is the process of dividing customers into groups based on common characteristics so companies can market to each group effectively and appropriately.

Problem statement

Avarto Financial Solutions, a Bertelsmann subsidiary, is a mail-order sales company in Germany interested in identifying segments of the general population to target with their marketing in order to acquire new clients more efficiently.

Datasets and inputs

Due to the confidentiality of the data, it is available only on the workspace provided by Udacity.

There are four data files and two description 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 marketing campaign targets; 42 982 persons (rows) x 367 (columns).
- Udacity_MAILOUT_052018_TEST.csv: Demographics data for individuals who were marketing campaign targets; 42 833 persons (rows) x 366 (columns).
- DIAS Information Levels - Attributes 2017.xlsx: is a top-level list of attributes and descriptions, organized by informational category.
- DIAS Attributes - Values 2017.xlsx: is a detailed mapping of data values for each feature in alphabetical order.

Each row of the demographics files represents a single person, but also includes information outside of individuals, including information about their household, building, and neighborhood.

The "CUSTOMERS" file contains three extra columns ('CUSTOMER_GROUP', 'ONLINE_PURCHASE', and 'PRODUCT_GROUP'), which provide broad information about the customers depicted in the file.

The original "MAILOUT" file included one additional column, "RESPONSE", which indicated whether or not each recipient became the company's customer. For the "TRAIN" subset, this column has been retained, but in the "TEST" subset it has been removed; it is against that withheld column that your final predictions will be assessed in the Kaggle competition.

Otherwise, all of the remaining columns are the same between the three data files.

Solution statement

1- Unsupervised learning:

   Clustering by using the information from the first two files to figure out how customers ("CUSTOMERS") are similar to or different from the general population at large ("AZDIAS")

2- Supervised learning:

   Making predictions on the other two files ("MAILOUT"), predicting which recipients are most likely to become a customer for the mail-order company.

Benchmark model

Submitting the results into the Kaggle competition and getting a score and seeing my rank.
https://www.kaggle.com/competitions/udacity-arvato-identify-customers/overview

In addition, I have access to the project layout for part 1 (unsupervised learning) provided by Udacity from another nanodegree. https://learn.udacity.com/nanodegrees/nd230-fwd-t4/parts/70a4cb52-2cce-4601-84a3-340ada73815d/lessons/30a5f851-823f-46ca-8ea7-50e1471194c9/concepts/bf44af3d-f908-4156-aed3-bd200be5f469

There are several blogs and implementations about this project that can be referenced for me, for example:
https://medium.com/@mt3915/customer-segmentation-for-arvato-bertelsmann-b0026efbb554
https://365datascience.com/tutorials/python-tutorials/pca-k-means/
https://medium.com/@tongxiaoling1022/create-a-customer-segmentation-report-for-arvato-financial-solutions-udacity-data-scientist-bb1194218e82
https://github.com/sallytxl/capstone
https://github.com/olgared/Capstone_Arvato_project_Term_2
https://github.com/sanjeevai/customer_segments_arvato/blob/master/README.md
https://github.com/sanjeevai/customer_segments_arvato/blob/master/Project_Rubric.pdf
https://github.com/patelatharva/Arvato_Customer_Segmentation
https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

Evaluation metrics

AUC (Area Under the Curve) for the ROC curve (Receiver Operating Characteristic curve), relative to the detection of customers from the mail campaign. 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).

The line plotted on these axes depicts the performance of an algorithm as we sweep across the entire output value range. We start by accepting no individuals as customers (thus giving a 0.0 TPR and FPR) then gradually increase the threshold for accepting customers until all individuals are accepted (thus giving a 1.0 TPR and FPR). The AUC, or area under the curve, summarizes the performance of the model. If a model does not discriminate between classes at all, its curve should be approximately a diagonal line from (0, 0) to (1, 1), earning a score of 0.5. A model that identifies most of the customers first, before starting to make errors, will see its curve start with a steep upward slope towards the upper-left corner before making a shallow slope towards the upper-right. The maximum score possible is 1.0, if all customers are perfectly captured by the model first.

Project design

1- Cleaning the data (the first two files)
2- Feature Transformation

3- Clustering
4- Train the model using the mailout training dataset.
5- Testing the model using the mailout testing dataset and submitting the results into Kaggle.

Resources:

1- https://www.shopify.com/blog/what-is-customer-segmentation
2- https://www.kaggle.com/competitions/udacity-arvato-identify-customers/overview
3- https://medium.com/@mt3915/customer-segmentation-for-arvato-bertelsmann-b0026efbb554
4- https://365datascience.com/tutorials/python-tutorials/pca-k-means/
5- https://medium.com/@tongxiaoling1022/create-a-customer-segmentation-report-for-arvato-financial-solutions-udacity-data-scientist-bb1194218e82
6- https://github.com/sallytxl/capstone
7- https://github.com/olgared/Capstone_Arvato_project_Term_2
8- https://github.com/sanjeevai/customer_segments_arvato/blob/master/README.md
9- https://github.com/sanjeevai/customer_segments_arvato/blob/master/Project_Rubric.pdf
10- https://github.com/patelatharva/Arvato_Customer_Segmentation
11- https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
12- https://learn.udacity.com/nanodegrees/nd230-fwd-t4/parts/70a4cb52-2cce-4601-84a3-340ada73815d/lessons/30a5f851-823f-46ca-8ea7-50e1471194c9/concepts/bf44af3d-f908-4156-aed3-bd200be5f469