Contents

Clustering

Problem domain/dataset and objectives

The dataset chosen for the clustering analysis is a publicly available CSV file downloaded from the UCI Machine Learning Repository. It relates to direct marketing activities run by a Portuguese bank, data from this dataset include information about the client's details, including attributes of age, job, marital status, level of education obtained, contact preference, etc.

```
Index(['age', 'job', 'marital', 'education', 'default', 'housing', 'loan',
       'contact', 'month', 'day_of_week', 'duration', 'campaign', 'pdays',
       'previous', 'poutcome', 'emp.var.rate', 'cons.price.idx',
       'cons.conf.idx', 'euribor3m', 'nr.employed', 'y'],
      dtype='object')
```

Fig.13. Presenting variable names - Python

The dataset consists of 41188 entries and 21 columns, feature extraction was required in data cleaning before the clustering analysis implementation.

The objective of the clustering analysis with the use of a machine learning clustering algorithm is to discover structures and patterns in high-dimensional data, as well as to group data with similar patterns together. This analysis aims to cluster Portuguese banks' customers based on their age, job, marital status, and education level. In marketing, customer segmentation is based on analysing customer details and finding patterns based on the clustering analysis.

Data Cleaning

First, examining data helped to present details of the dataset, for instance within the use of describe() function to gain knowledge about the central tendency, dispersion, and shape of a dataset's distribution.

Step.1. Subsetting a dataset

Data pre-processing began with subsetting the dataset, which allowed to decrease the number of variables to 4 for further analysis.

	age	job	marital	education
0	56	housemaid	married	basic.4y
1	57	services	married	high.school
2	37	services	married	high.school
3	40	admin.	married	basic.6y
4	56	services	married	high.school

Fig.24. Subsetting a dataset to 4 attributes

Step.2. Converting the age variable into a new categorical variable

Essentially, the categorical variable of age_bin is required for clustering analysis, therefore, converting the 'age' variable into a new 'age_bin' categorical variable was required.

	age	job	marital	education	age_bin
0	56	housemaid	married	basic.4y	50-60
1	57	services	married	high.school	50-60
2	37	services	married	high.school	30-40
3	40	admin.	married	basic.6y	30-40
4	56	services	married	high.school	50-60

Fig.25. Converting age into a new categorical variable

Step.3. Missing values

One of the crucial tasks when examining the data is ensuring that there are no missing values in the dataset. Pandas have a built-in function that allows detecting any missing values - isNull() After running the command with the addition of the sum() command, the terminal has an output that there are no missing values in the set. In addition, running nunique() allows counting the number of distinct elements in a specified axis.

Step.4. Converting categorical features into numerical features

As dataset attributes consist of strings, converting categorical features into numerical features was required as the third step of the pre-processing process. The reasoning behind this is that all input and output variables for machine learning techniques must be numeric. This implies that in order to fit and assess a model, categorical data must first be encoded into numerical.

	age	job	marital	education	age_bin
0	39	3	1	0	4
1	40	7	1	3	4
2	20	7	1	3	2
3	23	0	1	1	2
4	39	7	1	3	4

Fig.26. Converting categorical features into numerical features

Step.5. Dividing the dataset into labels and features

Data splitting is a crucial component in data science, especially when building machine learning models. This method aids in ensuring the accuracy of data model construction and the processes that use data models, such as machine learning.

```
#Data splitting
X = customer.drop('age_bin', axis=1) # Features
Y = customer['age_bin'] # Labels
print(X.shape)
print(Y.shape)
```

```
(41188, 4)
(41188,)
```

Fig.27. Data splitting. Dividing market_df into label and feature sets

Step.6. Normalising numerical features

A fundamental condition for many machine learning forecasting techniques is the standardization of a dataset. If the individual features do not more or less resemble standard normally distributed data, they may perform not efficiently. For instance, Gaussian with 0 mean and unit variance.

```
feature_scaler = StandardScaler()
X_scaled = feature_scaler.fit_transform(X)

X_scaled
```

```
array([[ 1.5334083 , -0.20157925, -0.2837415 , -1.75392459],
       [ 1.62938803,  0.91122681, -0.2837415 , -0.34973033],
       [-0.29020655,  0.91122681, -0.2837415 , -0.34973033],
       ...,
       [ 1.5334083 ,  0.35482378, -0.2837415 ,  1.05446393],
       [ 0.38165155,  1.46762984, -0.2837415 ,  0.58639918],
       [ 3.26104342,  0.35482378, -0.2837415 ,  0.58639918]])
```

Fig.28.Normalising numerical features

Step.7. Dimensionality Reduction using PCA

Principal Component Analysis (PCA) is a method for lowering the number of dimensions in a dataset while keeping the majority of the information. It makes use of the correlation between some dimensions and works to keep the number of variables to a minimum while maintaining the maximum degree of variation or knowledge about the distribution of the original data.

```
Variance explained by each of the n_components:  [0.36418117 0.27202165]
Total variance explained by the n_components:  0.636202818548712
```

PCA

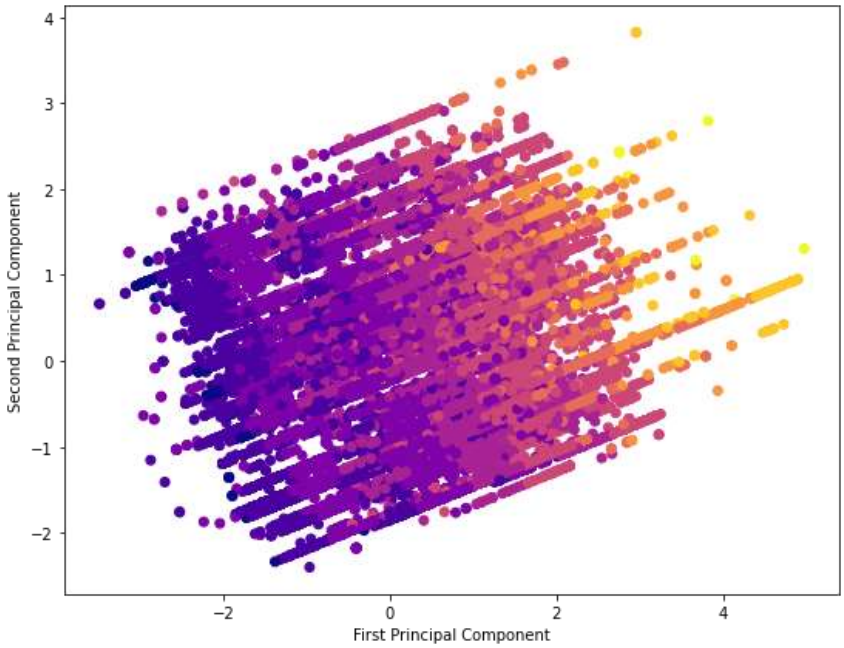


Fig.29. Principal Component Analysis

Elbow Plot Method

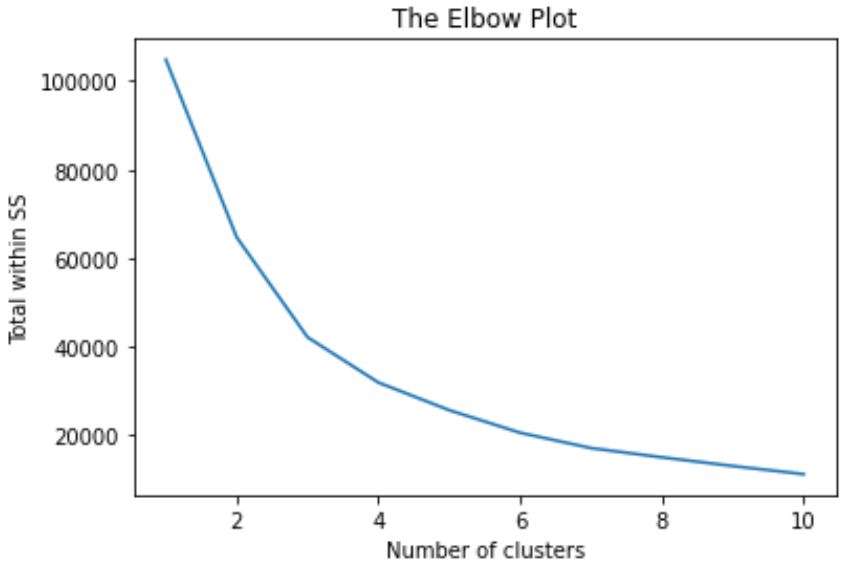
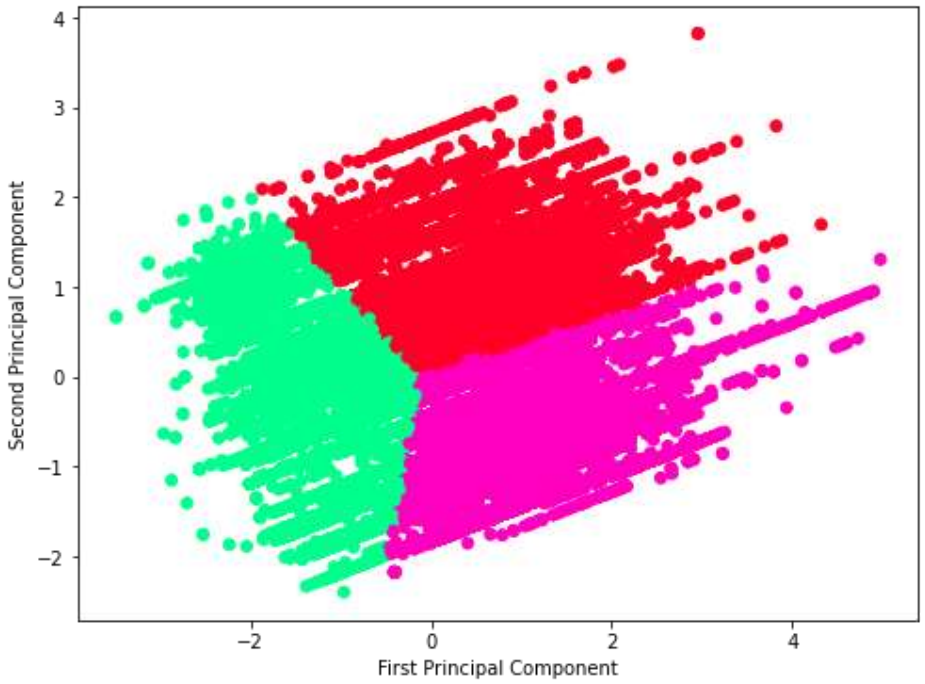


Fig.30.Establishing number of K clusters – line chart- Python

The Elbow Plot Method is used to find the number of K clusters. In the case of the line chart above the K=3

Visualizing clusters



Summary

Bousquet, Luxburg & Ratsch (2003) defined unsupervised learning as *"…. can be thought of as finding patterns in the data above and beyond what would be considered pure unstructured noise."* The unsupervised machine learning technique doesn't require a user to supervise the model, there is no training stage and it discovers information self-sufficiently. Marketers use this technique to find patterns across the customers. In this report, the dataset consisted of the client details, the patterns were able to be found accordingly.

Before applying the clustering technique, all the information in the dataset was converted to numeric so the mathematical distances could be calculated. All the steps taken through pre-processing and cleaning data, allowed for the implementation of clustering analysis afterwards.

Clustering analysis allowed to presentation of clusters as scatter plots, visualising them clearly and efficiently. There are some advantages and disadvantages of clustering, however, this method is used across several sectors, especially in marketing when customer segmentation is helpful during creating marketing campaigns. This type of machine learning technique is dynamic, efficient to implement plus can be implemented on big data. Nevertheless, predefining the number of clusters is not a straightforward decision for experienced data analysts.