Analyzing the Impact of Attributes on Term Deposit Subscriptions in Portuguese Banking Marketing Campaigns

UCI Bank Marketing Dataset

Dataset information -

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

Attributes -

17 Attributes in total - 7 Numeric and 10 Categorical

	Variable	Type
1	age	Numeric
2	job	Categorical
3	marital	Categorical
4	education	Categorical
5	default	Categorical
6	balance	Numeric
7	housing	Categorical
8	loan	Categorical
9	Contact	Categorical
10	month	Categorical
11	day	Numeric
12	duration	Numeric
13	campaign	Numeric
14	pdays	Numeric
15	previous	Numeric
16	poutcome	Categorical
17	y (target variable)	Categorical

OBJECTIVE - We aim to identify the attributes that have the highest impact on clients who are likely to subscribe to a term deposit.

PART 1 - EDA

Approach -

1. Explore Numeric Attributes
2. Explore Categorical Attributes
3. Remove Outliers - Numerical Attributes
4. Data Imputation - Categorical Attributes
5. Data Transformation and Normalization

EDA (Numeric Attributes) -

Mean of Numerical Attributes for Output Variable (y) -

y	age	balance	campaign	day	duration	pdays	previous
no	40.838	1303.715	2.846	15.892	221.183	36.421	0.502
yes	41.67	1804.268	2.141	15.158	537.295	68.703	1.17

From these averages, we can observe that clients who subscribed to a term deposit (represented by the "yes" category) tend to have certain characteristics compared to those who did not (represented by the "no" category):

1. Higher average balance: Clients with a higher average balance tend to have a higher rate of subscribing to a term deposit.

2. Longer average duration: Clients who had longer average contact durations during the campaign are more likely to subscribe to a term deposit.

3. More previous contacts: Clients who had a higher average number of previous contacts tend to have a higher likelihood of subscribing to a term deposit.

4. These three characteristics (higher balance, longer duration, and more previous contacts) are associated with a higher rate of subscribing to a term deposit.

EDA (Categorical Attributes) -

1. Subscription ratio of the number of clients who subscribe the product to the total number of clients in that specific **job category** :
   student (ratio=0.286)
   retired(ratio=0.227)
   unemployed(ratio=0.155)

2. While the majority of the clients are married, it seems that the clients who are **single** are the group that provides the highest subscribe rate among all
3. The clients who have **tertiary and unknown education** background are the groups that provides the highest subscribe rate among all.
4. The clients with **no hosing loan** are more willing to subscibe the product
5. The clients with **no personal loans** are more willing to subscibe the product
6. The clients who are **contacted by cell phone** are more willing to subscibe the product
7. May, July, and August are the top 3 months that the majority of clients are last contacted - according to bar plot
8. According to the quantitative table, March, December, and September have the highest rates for the clients to subscribe the product.
9. when looking at the yes/no distribution corresponded to the category, the success category, and the other category are the top 2 groups among all.

Remove Outliers - Numerical Attributes

Filter out potential outliers by removing rows where the values in specific columns deviate significantly from the central range defined by the IQR method. Calculated the 25th percentile (Q1) and 75th percentile (Q3) for the 'age', 'balance', 'day', 'duration', 'pdays', and 'previous' columns using the describe() method.

This outlier removal process helps ensure that the dataset is free from extreme values that could potentially skew the analysis or modeling results.

Data Imputation - Categorical Attributes

Performed a data imputation technique for categorical attributes. It focuses on three specific categorical attributes: "job", "education", and "contact".

By performing this data imputation technique, we replaced the "unknown" values in the specified categorical attributes with the mode category, effectively filling in missing or unknown values with the most frequent category for each attribute.

Data Transformation and Normalization

Transformed 10 categorical attributes into binary columns using one-hot encoding and normalized the data using MinMaxScaler. Categorical attributes are encoded using one-hot encoding, while 7 numeric attributes are normalized.

PART 2 - Customer Segmentation Analysis

Approach-

1. Feature Selection
2. K-Means Clustering
3. Principal Component Analysis (To visualise clusters)
4. Customer Segmentation Analysis

Feature Selection

Applied variance-based feature selection using *'VarianceThreshold'* to identify and retain only the features with variances greater than 0.1. Then displayed the selected features and their variances, as well as the features that are filtered out due to having variances below the threshold.

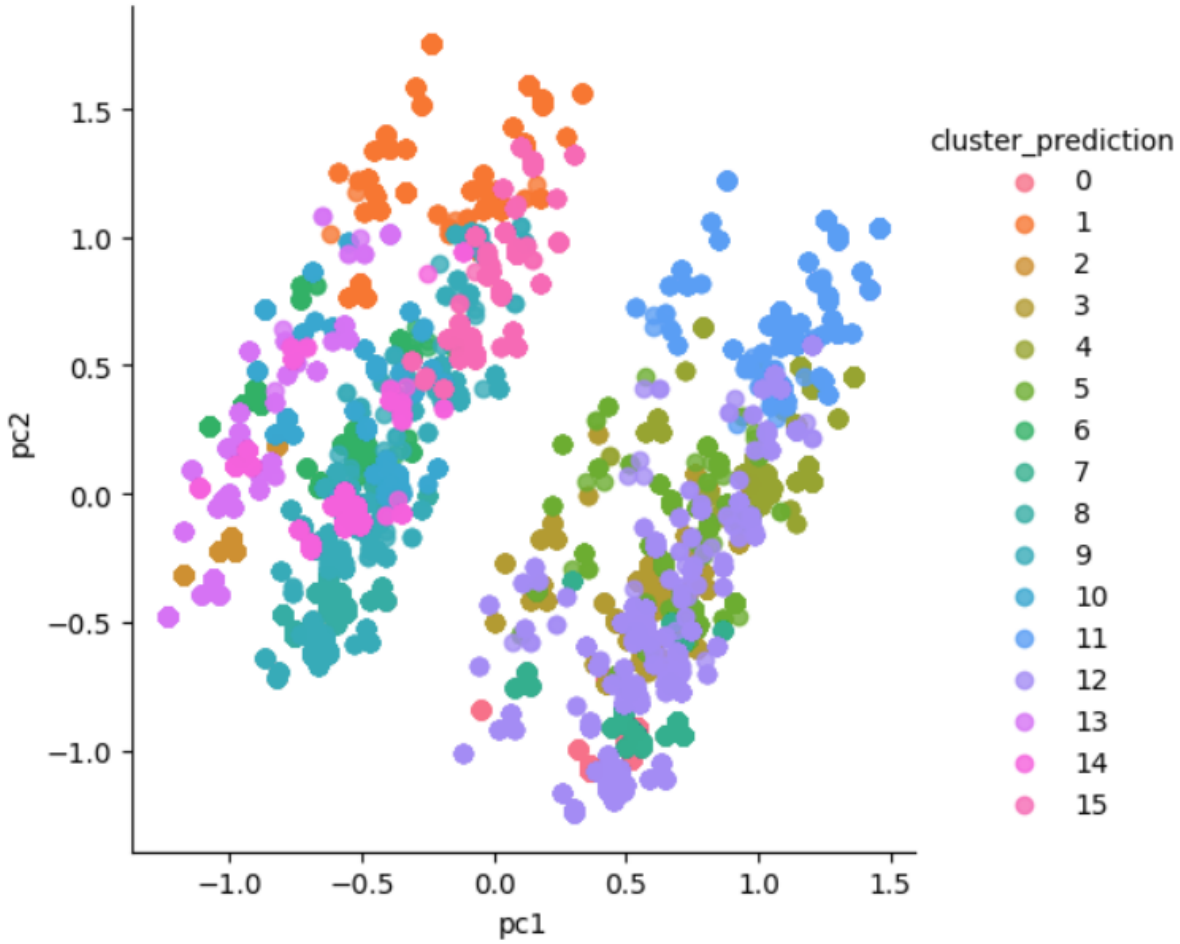
The features with variances below 0.1 have less variation in their values, suggesting they may not contribute significantly to the predictive power of the model. On the other hand, the features with variances above 0.1 have greater variation and are more likely to provide meaningful information for modeling and prediction tasks.

K-Means Clustering

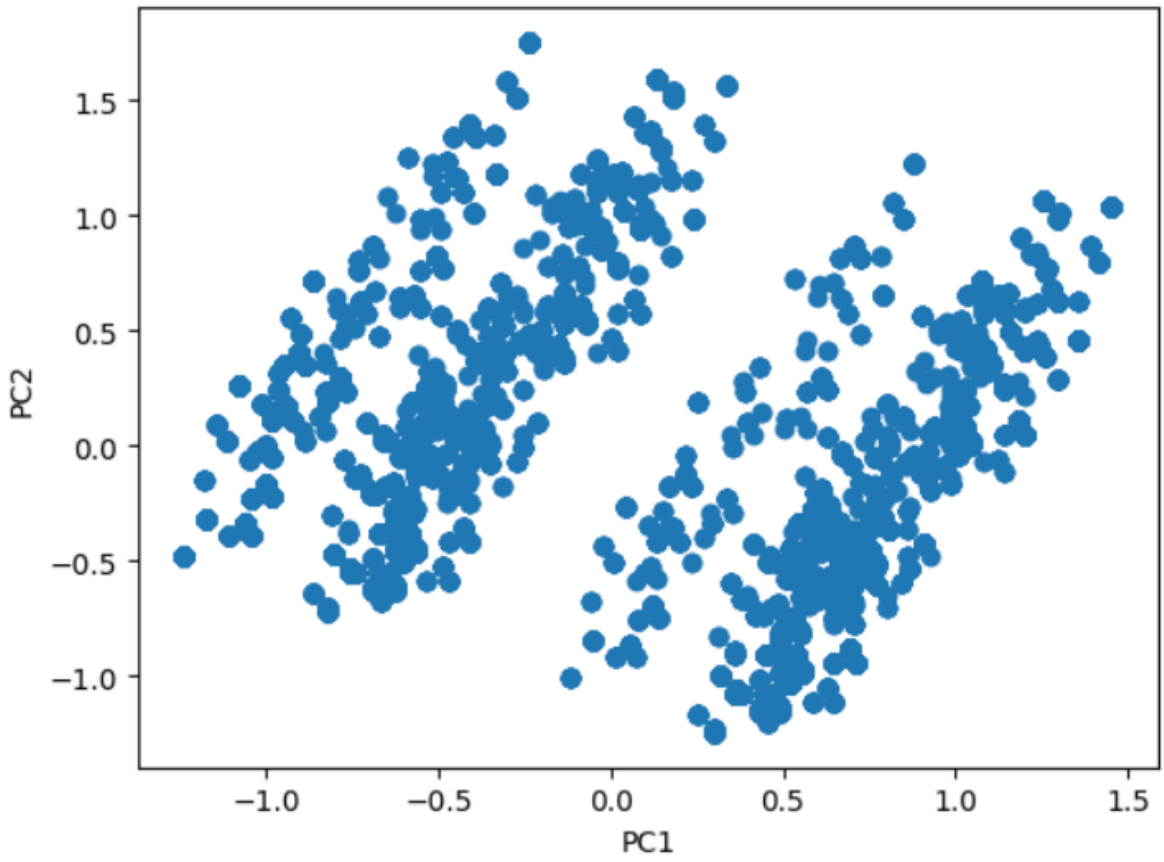
Performed K-means clustering on the dataset using different values of the number of clusters.
Calculated the inertia (intra-class similarity) and silhouette score for each number of clusters.
K-means aims to minimize the within-cluster sum of squared distances.
Silhouette score measures the quality of clustering by calculating the average distance between data points within clusters and the average distance between data points in different clusters.
Higher silhouette scores indicate better-defined clusters.
Elbow method plots the number of clusters against the inertia (intra-class similarity) or another metric and looks for the "elbow" point where the change in the metric significantly decreases.

PCA

Performed Principal Component Analysis (PCA) on the transformed training data to reduce its dimensionality to 2 components.



This Plot shows the outcome of K-means clustering with k=16. Based on the Silhouette score, when k=16, the score reaches 0.2509, which is the highest among the plotted values. However, the score's proximity to 0 suggests that there is significant overlap between clusters in higher dimensions, indicating limited distinctiveness among the clusters.



This plot illustrates the representation of the original data, consisting of 17 features, on a 2D plot. The data is transformed into a lower-dimensional space using two principal components, which individually capture 0.2 and 0.14 variance from the original data. This plot provides a visual understanding of the data distribution and its condensed representation in the reduced space.

Customer Segmentation Analysis

Approach-

a) Selected the clusters with the highest and lowest conversion rates: Cluster 7, Cluster 3, Cluster 8, and Cluster 1.

b) Next, analyzed the cluster centroids of these selected clusters to understand their representative characteristics.

c) To gain further insights, identified the top 10 features that significantly influence customers' decision to subscribe to the product using feature selection techniques. These features will be further analyzed for their impact on subscription rates.

1) ratio of the clients who subscribe the product to the total number of clients in the cluster

The top 4 clusters with their conversion rates are:

Cluster 7 - Conversion Rate: 0.032062

Cluster 3 - Conversion Rate: 0.029185

Cluster 8 - Conversion Rate: 0.027637

Cluster 1 - Conversion Rate: 0.026020

2) Cluster Representation

In **Cluster 7**, the highest ratio of customers who subscribed to the product (yes) is observed among those with secondary education (education_secondary), who are married (marital_married), have no loan (loan_no), and do not own a house (housing_no).

In **Cluster 3**, the highest "yes" ratio for subscribing to the product is associated with customers who have a marital status of single (marital_single). Other influential features include job position in administration (job_admin.), the month of August (month_aug), and the month of July (month_jul).

In **Cluster 8**, the most influential features for customers subscribing to the product are having no loan (loan_no), being married (marital_married), having secondary education (education_secondary), and having housing (housing_yes). The highest subscription ratio is observed for the feature "loan_no" with a ratio of 1.

In **Cluster 1**, the highest ratio of customers who subscribed to the product (yes ratio) is associated with having housing (housing_yes) and tertiary education (education_tertiary), while having a loan (loan_no) and being in a secondary education level (education_secondary) are less influential factors.

3) significant features that most influence the customers in the cluster to subscribe the product

In **Cluster 7**, the most significant features that influence the likelihood of customers subscribing to the product are the month of June, job position in administration, and job position in management.

In **Cluster 3**, the months of June and July, along with divorced customers, are the most influential factors in influencing the outcome variable 'y'.

In **Cluster 8**, the most influential factors for customers' decision to subscribe to the product are the month of August and job positions in management.

In **Cluster 1**, the months of July, August, and June, job position technician along with the married customers are influential factors in their decision to subscribe to the product.