مدينة زويل للعلوم والتكنولوجيا
Zewail City of Science and Technology

Machine Learning Project (Loan Dataset Analysis)

Problem definition:

Bank customers applying for a loan can provide a lot of information. Some of this information is useful and some isn't. The problem is which pieces of this information has a strong relation with the total loan amount and the loan duration so that the bank can use them to assess the loan.

Motivation

Understanding the data you have ,will make you make better decisions in different aspects. For our case ,understanding these loan related information should make us better decide who we should give loans to in the future and who should not.also ,would help us better target the people while marketing for the bank,which clients we should pay more attention to.So,it is clear how understanding these data would be beneficial to us.

Dataset Description:

We are provided with a row dataset of 11 columns, each holding a piece of information about a customer. They are as follows:

- Total O/S: Total loan amount
- TENOR_@Booking: Duration in months for the loan
- Loan Term: Duration in years
- Booking Date: The date on which the loan data is entered ● Maturity Date: the date on which a borrower's final loan payment is due. ● DPD: Days past dues, a metric that indicates whether a customer has been

  consistent in his/her repayments and if he/she has missed any, how many installments did the customer miss and by how many days. ● DOB: Date of Birth of the client
- Age of the client.
- Gender of the client.
- Customer Segment: classification of the customer based on the type of

  employment.

Approach and methodology:

- Sarting by doing the necessary preprocessing :

1)Cleaning the data from null and duplicate values (since nan values are not too much) we dropped it.

2)Encoding the categorical variables to be able to deal with the data set. Such encoding like : encoding Gender ,customer segmentation and for dates in our dataset we used unix timestamp to represent them

- Data Visualization :To examine the relationship between the different attributes and the loan duration and total loan amount we used the following univariate and multivariate visualization techniques:
    - Heat maps , histograms and count plots
    - PCA analysis technique (multivariate visualization)

  From this data analysis ,we saw some relation between the different attributes in the dataset and understood distribution of the data.

  Some Data analysis result :

    - Heatmap results:
        - Loan term is highly correlated with tenor_@Booking ,maturity date
        - Total amount is most correlated with customer segmentation
        - Gender is least correlated with all attributes
        - TENOR@BOOKING least correlated with all other attributes
        - AGE AT MATURITY is highly correlated with age (makes sense) Etc

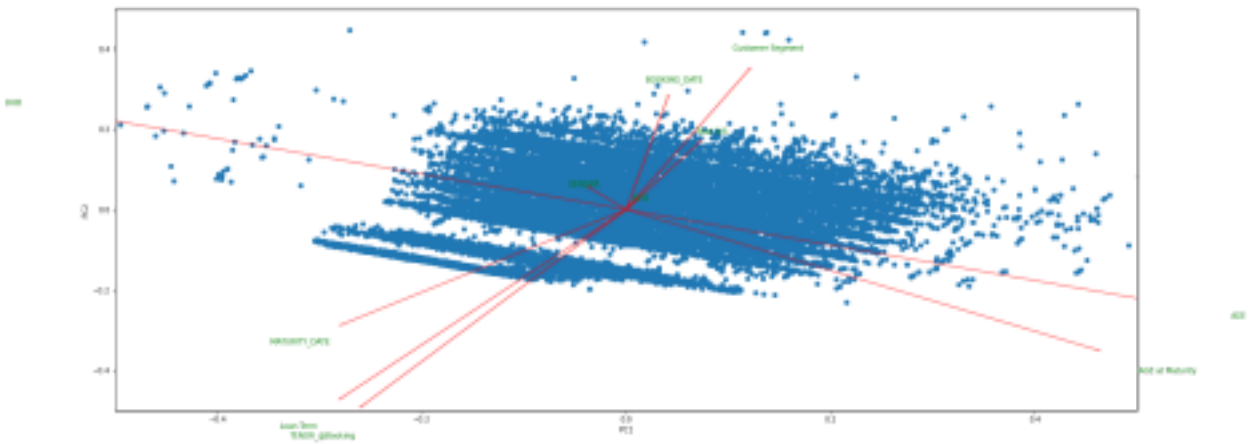
          ……
    - Other findings

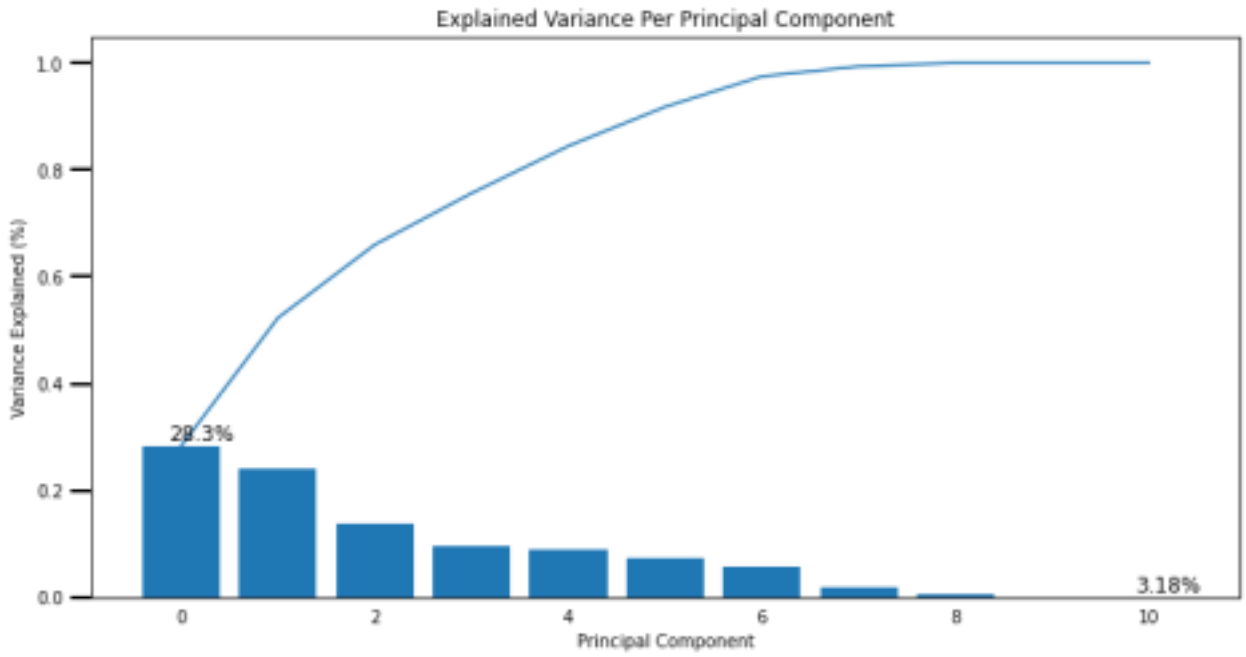
- Age of the Most of clients between 29 :48
- Males representing 83.97 ,females 16.03

- Performing pca analysis on our dataset :

1)To further examine the relations between the different attributes in our dataset 2)to use the generated pcas in our clustering techniques.

PCA results in 10 principal components with most of the variance in the first 5 ones.





- PCA analysis and results (for first two components) :

  1)Gender and DBD attributes are less important than other ones

  2)The first principle component (first axis) provides information about age ,age at maturity and date of birth.

  3)second PC does not provide such more information or explain high variance

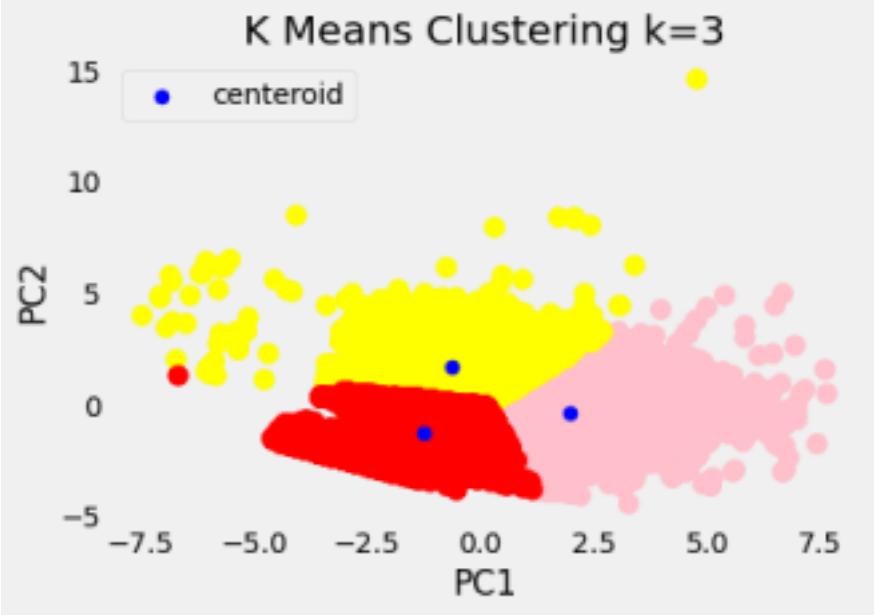
- K-means clustering :

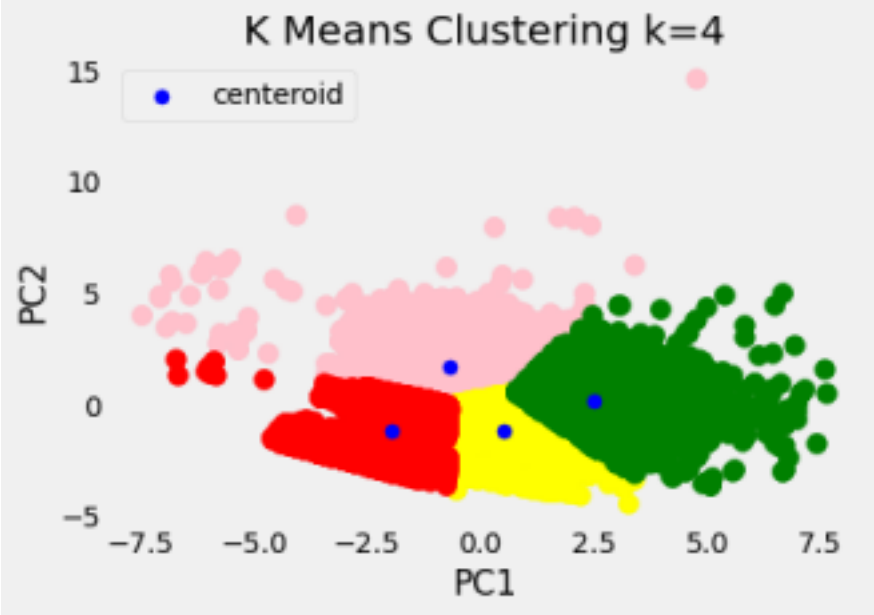
  Our data now consists of the first 2 PC (principle component)

  To decide the right k ,we used the elbow method to suggest some k to test

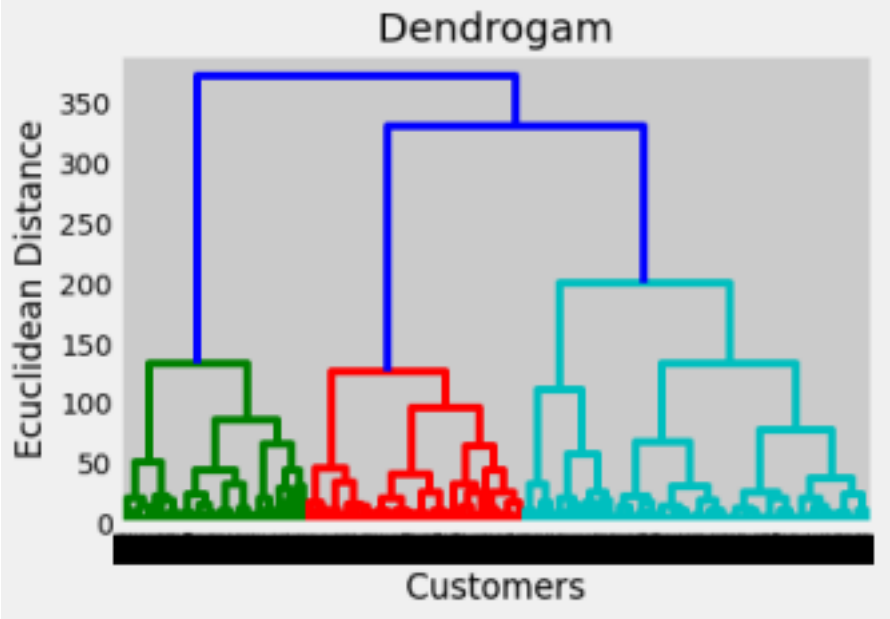
  Next plot is showing the elbow

- From the above plot : It suggests k o start from 3 : 7 (it does not provide such information)
- We fitted our model with k =3 ,4 , 6,7

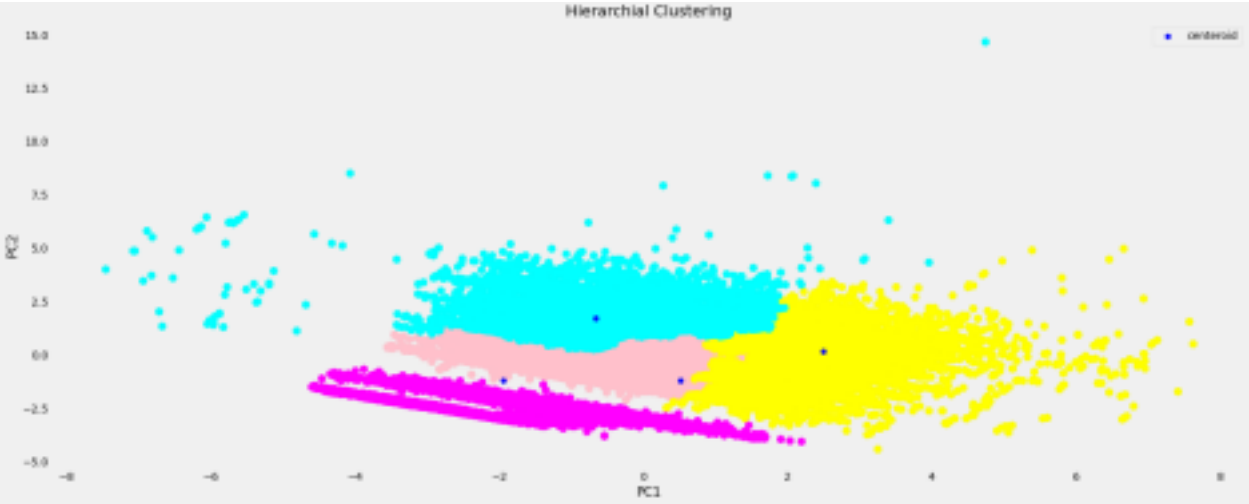




- Hierarchical clustering
    - Using dendrogram to suggest values for k



- Dendrogram suggests 3,4 but other values like 6 also possible
- Resultingclusterfrom



Evaluation of different models:

1. To evaluate our model ,we used silhouette score which is a metric to evaluate how good the clusters are through values ranges from -1 :1 such that:
    - 1 means excellent clustering with all the clusters are apart from each other
    - -1 means worst clusters
    - Ranges from 1 : -1 from excellent to bad clusters

2. silhouette scores for our model

K for data with 2 principle components silhouette scores
3 .373

4 .348
7 .353

Based on the above scores clustering the data into 3 clusters gives the best result.

K for data with 2 principle components silhouette scores
k=3 .353

Libraries used on :

Library	Usage
matplotlib	Pyplot figure plotting
seaborn	**Data visualization**
numpy	**Multidimensional array processing**

pandas Data analysis

random generate random numbers

KMeans Apply k-means

scipy.cluster.hierarchy Hierarchical clustering

Conclusion :

From clusters we get ,there exist pattern in the data on which we can
1) build a classification model that classify to each cluster

2)from the univariate visualization , there exist correlation between different attributes for example

- We can build a regression models that could expect loan term ●

  Other regression models based on correlation between the different attributes

3)Build a model to predict the total amount of the loan but we need another attribute which is income