Credit Card Clustering and Segmentation

Unsupervised Learning

Domain: Banking, Finance

Business Context:

This case requires to develop a customer segmentation to define marketing strategy. The sample Dataset summarizes the usage behavior of about 9000 active credit card holders during the last 6 months. The file is at a customer level with 18 behavioral variables.

Data Dictionary:

CUSTID: Identification of Credit Card holder (Categorical)
BALANCE: Balance amount left in their account to make purchases (
BALANCEFREQUENCY: How frequently the Balance is updated, score between 0 and 1 (1 = frequently updated, 0 = not frequently updated)
PURCHASES: Amount of purchases made from account
ONEOFFPURCHASES: Maximum purchase amount done in one-go
INSTALLMENTSPURCHASES: Amount of purchase done in installment
CASHADVANCE: Cash in advance given by the user
PURCHASESFREQUENCY: How frequently the Purchases are being made, score between 0 and 1 (1 = frequently purchased, 0 = not frequently purchased)
ONEOFFPURCHASESFREQUENCY: How frequently Purchases are happening in one-go (1 = frequently purchased, 0 = not frequently purchased)
PURCHASESINSTALLMENTSFREQUENCY: How frequently purchases in installments are being done (1 = frequently done, 0 = not frequently done)
CASHADVANCEFREQUENCY: How frequently the cash in advance being paid
CASHADVANCETRX: Number of Transactions made with "Cash in Advance"
PURCHASESTRX: Numbe of purchase transactions made
CREDITLIMIT: Limit of Credit Card for user
PAYMENTS: Amount of Payment done by user
MINIMUM_PAYMENTS: Minimum amount of payments made by user
PRCFULLPAYMENT: Percent of full payment paid by user
TENURE: Tenure of credit card service for user

Steps:

1. Preprocessing the data (15 points)
    a. Check a few observations and get familiar with the data. (1 points)
    b. Check the size and info of the data set. (2 points)
    c. Check for missing values. Impute the missing values if there is any. (2 points)

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.

d. Drop unnecessary columns. (2 points)
e. Check correlation among features and comment your findings. (3 points)
f. Check distribution of features and comment your findings. (3 points)
g. Standardize the data using appropriate methods. (2 points)

2. Build a k-means algorithm for clustering credit card data. Kindly follow the below steps and answer the following. (10 points)
    a. Build k means model on various k values and plot the inertia against various k values.
    b. Evaluate the model using Silhouette coefficient
    c. Plot an elbow plot to find the optimal value of k
    d. Which k value gives the best result?

3. Apply PCA to the dataset and perform all steps from Q2 on the new features generated using PCA. (15 points)

4. Create a new column as a cluster label in the original data frame and perform cluster analysis. Check the correlation of cluster labels with various features and mention your inferences. (Hint - Does cluster 1 have a high credit limit?) (5 points)

5. Comment your findings and inferences and compare the performance. Does applying PCA give a better result in comparison to earlier? (5 points)

Learning Outcome:

- Clustering
- K-means
- PCA

Source: https://www.kaggle.com/arjunbhasin2013/ccdata

Proprietary content. © Great Learning. All Rights Reserved. Unauthorized use or distribution prohibited.