Project Proposal

Cohort B Team 7 - Kunpeng Huang, Yoki Liu, Lyufan Pan, Yunlei Zhou, Sherry Zuo

Problem that motivates the analysis:

The problem that motivates us is the segmentation of customers that could be very necessary and powerful to define marketing strategies. The credit card usage behavior of customers with 18 behavioral features seems to be a perfect dataset for us to explore the market segmentation. Besides, the problem of credit card fraud would be a potentially meaningful research direction for our work.

Dataset:

Credit Card Dataset for Clustering
https://www.kaggle.com/arjunbhasin2013/ccdata

The sample Dataset summarizes the usage behavior of 8950 active credit cardholders during the last 6 months. The file is at a customer level with 18 behavioral variables.

Following is the Data Dictionary for Credit Card dataset :

- CUST_ID: Identification of Credit Cardholder (Categorical)
- BALANCE: Balance amount left in their account to make purchases
- BALANCE_FREQUENCY: How frequently the Balance is updated, score between 0 and 1 (1 = frequently updated, 0 = not frequently updated)
- PURCHASES: Amount of purchases made from the account
- ONEOFF_PURCHASES: Maximum purchase amount did in one-go
- INSTALLMENTS_PURCHASES: Amount of purchase done in installment
- CASH_ADVANCE: Cash in advance given by the user
- PURCHASES_FREQUENCY: How frequently the Purchases are being made, score between 0 and 1 (1 = frequently purchased, 0 = not frequently purchased)
- ONEOFF_PURCHASES_FREQUENCY: How frequently Purchases are happening in one-go (1 = frequently purchased, 0 = not frequently purchased)
- PURCHASES_INSTALLMENTS_FREQUENCY: How frequently purchases in installments are being done (1 = frequently done, 0 = not frequently done)
- CASH_ADVANCE_FREQUENCY: How frequently the cash in advance being paid
- CASH_ADVANCE_TRX: Number of Transactions made with "Cash in Advance"
- PURCHASES_TRX: Number of purchase transactions made

- CREDIT_LIMIT: Limit of Credit Card for user
- PAYMENTS: Amount of Payment done by the user
- MINIMUM_PAYMENTS: Minimum amount of payments made by the user
- PRC_FULL_PAYMENT: Percent of full payment paid by the user
- TENURE: Tenure of credit card service for user

Proposed analysis methodology:

1. Clean the dataset:
    a. Check and replace missing values in the dataset
    b. Adjust different value types
    c. Chose meaningful variables for analysis
    d. Detect abnormal values and manipulate
2. Preview of Data
    a. Summary Stats
    b. Describing the Data
        i. Variable Names and explanation
        ii. Visualization
3. Build different models and tune the models:
    a. Logistic Regression
    b. Stepwise Regression: forward, backward
    c. Penalized Regression: Lasso, Ridge
    d. Decision Tree
    e. k-means
    f. PCA
    g. EFA
    h. t-SNE
    i. Other
4. Find the best model and evaluate the performance.
    a. Find the best model by comparing the accuracy such as precision, recall, F1
    b. Proceed the best model to seek the optimal results