**Project Proposal: Team 6**

Team Members: Xiaoqi Hu, Fernanda Lin, Haolan Ma, Chenhang Niu, Yi Yu

Date: 2019-11-3

## Problem Statement

Over time, credit card companies have received support data about the various customers they own. The company collects data on facts related to customers, such as their balances, purchases, cash advances, credit lines, and more. Our team's mission is to get meaningful insights from the data and then develop strategies that the company can target customers and increase the amount of credit card sales, which in turn increases revenue. The data set used in the analysis consists primarily of behavioral and non-label data related to credit card transactions. The main goal is to implement clustering to indicate the customer base that best fits the data. Based on the existence of high-dimensional features and related variables of the dataset, we will compare the clustering results based on the original data with the results obtained by implementing the correlation algorithm to reduce the dimension and correlation.

## Data set Introduction

## The dataset[[1]](#footnote-0) summarizes the usage behavior of 8950 active credit card holders within 6 months period. It has 8950 rows(observations) and 18 columns(variables). There is 1 NA in CREDIT\_LIMIT, and 313 NAs in MINIMUM\_PAYMENTS.

### Exhibit 1: Data Dictionary

|  |  |  |
| --- | --- | --- |
| **Variable Names** | **Explanation** | **Data Format** |
| CUST\_ID | Identification of Credit Card holder | (character) |
| BALANCE | Balance amount left in their credit card account to make purchases | (numeric) |
| BALANCE\_FREQUENCY | How frequently the Balance is updated, score between 0 and 1 (1 = frequently updated, 0 = not frequently updated) | (numeric) |
| PURCHASES | Amount of purchases made from account | (numeric) |
| ONEOFF\_PURCHASES | Maximum purchase amount done in one-go | (numeric) |
| INSTALLMENTS\_PURCHASES | Amount of purchase done in installment | (numeric) |
| CASH\_ADVANCE | Cash in advance given by the user | (numeric) |
| PURCHASES\_FREQUENCY | How frequently the Purchases are being made, score between 0 and 1 (1 = frequently purchased, 0 = not frequently purchased) | (numeric) |
| ONEOFF\_PURCHASES\_FREQUENCY | How frequently Purchases are happening in one-go (1 = frequently purchased, 0 = not frequently purchased) | (numeric) |
| PURCHASES\_INSTALLMENTS\_FREQUENCY | How frequently purchases in installments are being done (1 = frequently done, 0 = not frequently done) | (numeric) |
| CASH\_ADVANCE\_FREQUENCY | How frequently the cash in advance being paid | (numeric) |
| CASH\_ADVANCE\_TRX | Number of Transactions made with "Cash in Advance" | (numeric) |
| PURCHASES\_TRX | Number of purchase transactions made | (numeric) |
| CREDIT\_LIMIT | Limit of Credit Card for user | (numeric) |
| PAYMENTS | Amount of Payment done by user | (numeric) |
| MINIMUM\_PAYMENTS | Minimum amount of payments made by user | (numeric) |
| PRC\_FULL\_PAYMENT | Percent of full payment paid by user | (numeric) |
| TENURE | Tenure of credit card service for user | (numeric) |

## Analysis Methodology

## *Data Cleaning*

First of all, every data analysis involves the process of data cleaning. As for our dataset, we have missing values in “MINIMUM\_PAYMENTS” and “CREDIT\_LIMIT”, and we are going to replace these missing values with their means. In addition, we will format all the variables into lower cases and replace spaces/commas with underscores. The final step in data cleaning is to split our data into 80% training data set and 20% test data set.

1. *Exploratory Analysis*

To better investigate the relationship between variables, identify patterns, and present to our audiences, we plan to do some exploratory analysis by using ggplot, correlation matrix. The final charts, plots, and correlation map will be selectively included in the final presentation and report.

1. *Modeling*

To find the best model for this analysis our team decided to experiment with different models and find the ones which will provide us with the most accurate predictions. As of now, we will try using some of the unsupervised machine learning methods we have explored in class, such as k-means clustering, hierarchical clustering, etc. In short, we will implement Clustering/Segmentation Methods and Dimensionality Reduction Techniques. Elbow criterion method and silhouette coefficient method will be applied to identify the number of centers while Eigenvalue will be used to identify key principal components.

Last but not least, as for challenges, the main challenge we may face is determining the best model for our target problem, which is the most accurate model. The dataset we are going to use contains less than 10,000 observations. In this case, our result for each model might look very similar to others so that it would be harder for us to pick the best-working one.

1. *Data Source:*[*https://www.kaggle.com/arjunbhasin2013/ccdata*](https://www.kaggle.com/arjunbhasin2013/ccdata) [↑](#footnote-ref-0)