# Classification and Attributes of Businesses Using [redacted]

Matheus Schossler

March 14, 2022

## Introduction

Increasing the [redacted] is [redacted]'s mission, and its [redacted] are used by businesses of every size. Thus, understanding [redacted] customers' payment activity data can be challenging but it is a robust guide to improve [redacted]'s products and specialized services. Building the [redacted] successfully must be data-driven and evidence-based. Here we link payment activity data with relevant information about [redacted] customers to inform [redacted]'s business strategies.

In this report, we identify general attributes of businesses' payment activity with [redacted] based on their transaction history. The results in this report could be used to help [redacted] better understand customer behavior, thus guiding the decision-making regarding personalized customer experiences. We first analyze the payment activity history of merchants processing with [redacted] between January 1st, 2033, and December, 31st, 2034. Then, we classify [redacted]'s customers based on their data and generate attributes for each of them, including predicting near-future transaction activity behaviors.

We optimally split merchants into 5 classes employing the K-means clustering algorithm. Each class has attributes that help us understand the kind of businesses processing with [redacted]. For example, we can point out customers that are more likely to do transactions frequently and customers that are more likely to stop using [redacted] in the months following 2034. This was found by comparing their attributes with the attributes of other classes of businesses in the dataset.

## Features and Preprocessing of Data

We are first interested in building a set of features that allow us to infer the attributes and predict the behavior of merchants. The dataset available contains each transaction for the years 2033 and 2034, with the merchant name and transaction amounts. Grouping the data by merchant enables us to compare different customers. There are 14,351 unique merchants in the dataset with a high variance range in the number of transactions and amounts, as shown in Table 1.

The transaction activity data can be used to create a time series with the transaction amounts and the number of transactions for each merchant. Therefore, the set of features for each merchant is the amount of money and the number of transactions per unit of time. This time unit was chosen as 1 week because this is a good time range to understand the transaction behaviors spanning over two years. This choice decreases the noise created by variation in day-to-day activity but does not destroy much of the information about the true time of the transactions.

Most merchants did not have transactions throughout the full 2 years period, but their transactions spanned a smaller range of time. This happens because merchants started and/or stopped processing someday in 2033 or 2034. To compare a merchant to another merchant we have extended their time series for the 2 full years filling the extra days with zero transactions processed. Figure 1 shows in a graphical representation an example of a preprocessed time series for the weekly transaction amounts by the merchant '005e8bb6fb'.

The time series for the total weekly amount of processed money per merchant is found to have a high correlation with the time series of the weekly number of transactions per merchant. This is expected because weeks with a large number of transactions indicate a large amount of money transacted by a given merchant. Therefore we have

| | Average merchant total amount | Average merchant number of transactions |
|---|---|---|
| **mean** | 16'333 | 105 |
| **std** | 64'317 | 528 |
| **min** | 2 | 1 |
| **25%** | 363 | 3 |
| **50%** | 1'603 | 11 |
| **75%** | 8'231 | 45 |
| **max** | 2'369'072 | 25512 |

Table 1: Description of the total transactions amount ($) and the total number of transactions per merchant.

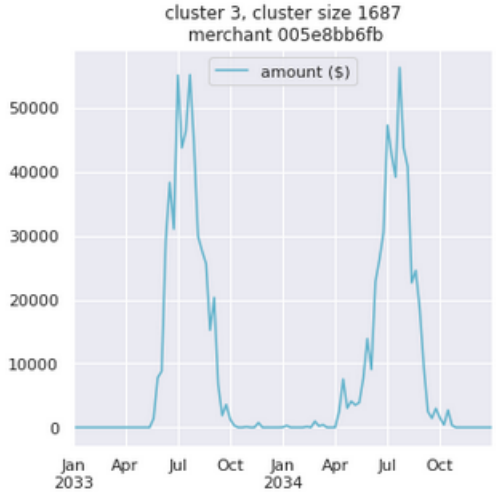



Figure 1: Amount transaction activity over 2 years for the merchant '005e8bb6fb'. Time series example.

chosen to not include the time series for the number of transactions in the final version of our model.

## Answering the Prompt: Model and the Results

The prompt is to find attributes for the [redacted] customers in the dataset, then predict the most likely merchants to stop processing with [redacted]. To achieve this task we believe is best to classify the merchants in groups or classes with somehow similar transaction activity behavior. However, some classes are more similar to each other in comparison to others. Comparing these classes allow us to predict the near future behavior of one class *if* it is similar to the past transaction activity behavior of another class. In the next subsection, we will explain our classification model, then show our results.

### The Clustering Algorithm[1]

The classification problem at hand is to classify merchants with similar transaction activity behavior in the unlabeled dataset. We use an unsupervised clustering algorithm, the K-means algorithm, to find these clusters of merchants with similar transaction activity.

Scaling the transaction activity for each merchant is essential to better compare businesses of different sizes. We choose the maximum absolute scaling to scale the transaction amounts per week per merchant between 0 and 1 because we want to have a comparable range between different merchants. In this scaling scheme, the behavior

[1]Check file transaction_history_analysis.ipynb for more technical details. However, the class ID may be different from this report due to the randomness of the centroid seeds of the K-means algorithm.

of the transaction activity over time is the most relevant feature for the classifier.

The K-means algorithm was utilized from the sklearn library. We first used the elbow method to find the best number of clusters that can classify well without overfitting. We find that 5 clusters are an optimal number to find classes of merchants with similar payment activity but not splitting them into too many different classes only having minor dissimilarities. The number of times the K-means runs with different centroid seeds had to be adjusted from the default in sklearn, 10, to 200 to make sure the algorithm finds the global minimum of the within-cluster sum of squares (WCSS) function. This relatively high number of runs is probably needed because of the large variance in the dataset.

## The Results

The classification results considerably reduced the variance, within each class, of the total transaction amount processed per merchant and the total number of transactions per merchant, in comparison to the unclassified results of Table 1. The average results for each class are summarized in Table 2.

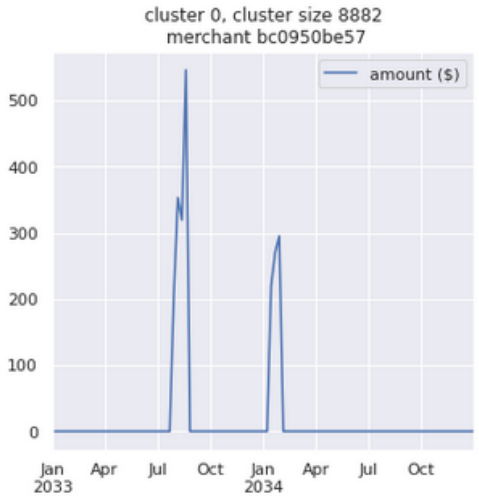
Each class has samples with average attributes that allow us to characterize the merchants in it. Limitations to this classification are discussed below. Class 1 has the smallest size, with less than 400 merchants. It is also the class that makes the most number of transactions with the largest average total amount of money processed per merchant with [redacted]. Businesses in this class used [redacted] for an average time interval of approximately 515 days over 2 years. In other words, the merchants in this class have a high processing frequency with [redacted], as can be seen in the sample example in Figure 2 (b). Class 2 has similar attributes as class 1, but merchants in this class have started using [redacted] near the end of the 2 years: an average of 9 months after the merchants of class 1 started. Analyzing the subsequent years would allow confirmation that class 2 could be merged with class 1 instead of class 3.

Class 3 has merchants that processed with [redacted] for a similar time interval as merchants in class 1: approximately 420 days on average. The difference, however, is the substantially smaller average total transaction amount per merchant and the average number of transactions per merchant. These merchants, on average, do business seasonally with a defined period for their payments activity as shown in the examples of Figures 1 and 2 (d).

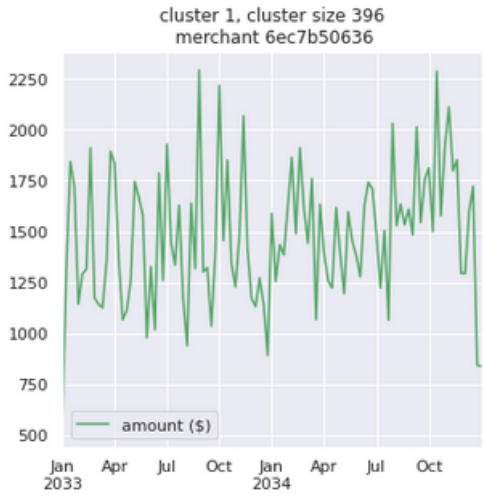
| **Class ID** | **# of merchants in the class** | **Average total amount** | **Average number of transactions** | **Average first transaction** | **Average last transaction** | **Average days w/ [redacted]** |
|---|---|---|---|---|---|---|
| 1 | 400 | $115'000 | 1240 | 2033-07-31 | 2034-12-26 | 515 |
| 2 | 940 | $38'000 | 275 | 2034-03-26 | 2034-12-24 | 275 |
| 3 | 1680 | $43'000 | 215 | 2033-09-28 | 2034-11-19 | 420 |
| 0 | 8880 | $6'600 | 32 | 2034-01-07 | 2034-06-05 | 150 |
| 4 | 2450 | $9'400 | 50 | 2034-07-16 | 2034-12-18 | 155 |

Table 2: General attributes of the average merchant in each class.

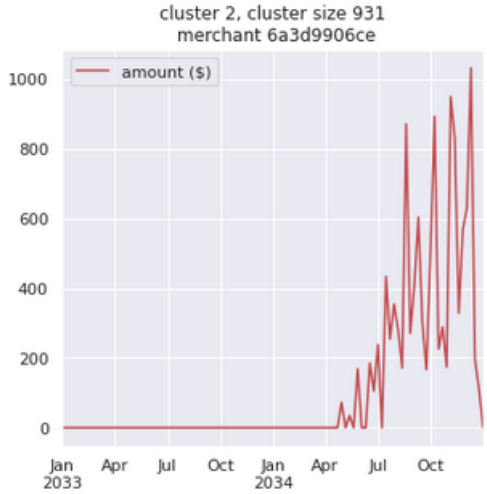
Businesses that have stopped processing with [redacted] earlier on in the 2 years, the churns, are in the largest class; class ID 0. One of the samples is plotted in Figure 2 (a). It is observed that the merchants in this class use [redacted] fewer times and less frequently than merchants from classes 1, 2, or 3. Also, it is noticed that merchants in this class make their last transaction, on average, about 6 months before their counterparts in other classes. [redacted] has low retention of merchants in this class. We notice that class 4 has the second-lowest average total amount of money transacted as well as the total number of transactions per merchant. We believe these are merchants that are most likely becoming churns in the near future.



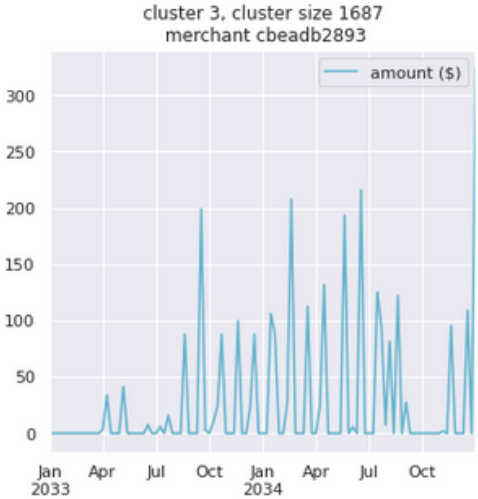

(a) Class 0 sample



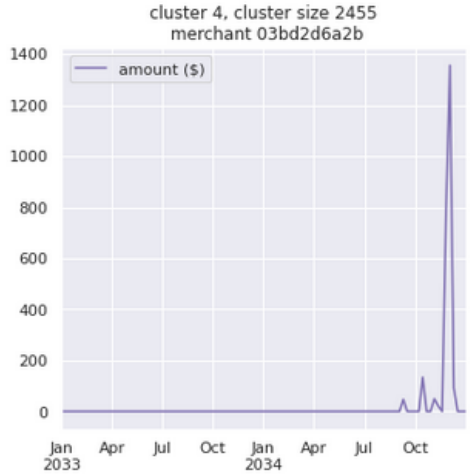

(b) Class 1 sample




(c) Class 2 sample




(d) Class 3 sample




(e) Class 4 sample

Figure 2: Time series sample from each cluster.

### Results Robustness

The identification of merchants based on the average attributes for each class works properly for many of them, for many others, however, it is not so accurate. The examples where this classification was not found to be accurate are located at far distances from the cluster centroids of the classification algorithm. For these merchants, our model does not always assign them to the classes that one would anticipate by visualizing their transaction activity time series. This classification could be improved by using more involved models, such as unsupervised neural networks, or supervised neural networks with a human-labeled dataset. However, we believe that more data for longer periods and over more merchant characteristics might help us understand better their activity and likelihood to stop processing with [redacted].

## Conclusion

In this report, we analyze the transaction activity data from [redacted] customers for the 2 years from January 1st, 2033, to December 31st, 2034. We model the attributes of merchants in this dataset according to their transaction activity behavior as a time series. We use the K-means clustering algorithm to cluster merchants into 5 classes and extract their average attributes and identify the types of businesses.

We find that class 1 of merchants constituted, on average, by high frequent users. Class 2 is composed of high frequent users, but the merchants in this class started using [redacted] later on in the dataset period. The transaction activity after 2034 must be analyzed to confirm whether this class can be combined with class 1. See Table 2 for the attributes of classes 1 and 2.

Merchants in class 3 have consistently used [redacted] for approximately 420 days, but less frequently than class 1 and 2. These are the seasonal users. The average total amount of money these merchants processed with [redacted] was around $43'000.

Class 0 is formed, on average, by merchants with significantly lesser total dollar amounts and fewer transactions than classes 1, 2, and 3. Many of these merchants were no longer actively using [redacted] by the end of 2034. Their last transaction happened on June 5th, 2034, on average. These merchants have churned. [redacted] has low retention of customers in this class. The merchants in class 4 are the most likely to churn in the near future. They are also low frequent users, with lesser total amounts of money processed with [redacted] than classes 1, 2, and 3. Their average first transaction happened on July 16th, 2034, more than a half year later than merchants in class 0, thus data for later years could confirm whether the merchants in this class churned or not. These are customers that need extra attention from [redacted] to increase their retention in the near future.

The data-driven evidence found in this report could be used to help better understand customer behavior, thus guiding the strategies and decision-making at [redacted]. In particular, analyses of additional class 0 characteristics could better elucidate the underlying cause of churn and inform the development of strategies to boost retention from the almost 2,500 customers in class 4.