ANALYZING INFLUENTIAL FACTORS ON THE RECOVERY TIME OF NON-PERFORMING LOANS: A TIME SERIES AND MACHINE LEARNING APPROACH

Vahrey Sitsuksai

A Thesis Submitted in Partial
Fulfillment of the Requirements for the Degree of
Master of Science (Business Analytics and Data Science)
School of Applied Statistics
National Institute of Development Administration
2023

ANALYZING INFLUENTIAL FACTORS ON THE RECOVERY TIME OF NON-PERFORMING LOANS: A TIME SERIES AND MACHINE LEARNING APPROACH

Vahrey Sitsuksai
School of Applied Statistics

..... Major Advisor
(Assistant Professor Ekarat Rattagan, Ph.D.)

The Examining Committee Approved This Thesis Submitted in Partial Fulfillment of Requirements for the Degree of Master of Science (Business Analytics and Data Science).

..... Committee Chairperson
(Assistant Professor Taweesak Samanchuen, Ph.D.)

..... Committee
(Assistant Professor Ekarat Rattagan, Ph.D.)

..... Committee
(Associate Professor Surapong Auwatanamongkol, Ph.D.)

ABSTRACT

Title of Thesis	ANALYZING INFLUENTIAL FACTORS ON THE RECOVERY TIME OF NON-PERFORMING LOANS: A TIME SERIES AND MACHINE LEARNING APPROACH
Author	Miss Vahrey Sitsuksai
Degree	Master of Science (Business Analytics and Data Science)
Year	2023

Non-Performing Loans (NPLs) are critical factors that impede economic growth. Efficient management systems are required to expedite the resolution process for borrowers and reduce the recovery time for the Asset Management Company (AMC). However, managing NPLs remains a challenge due to their complex behavior, which is difficult to understand. Consequently, this paper aims to enhance the understanding of borrower behavior and characteristics that affect recovery time, thereby enabling more effective loan recovery strategies. In this paper, we propose a combination of time-series clustering using Dynamic Time Warping (DTW) and random forest classification to analyze the impact of various features on the clustering results of loan recovery time based on collection patterns in the context of 2,839 loans. Our findings reveal that borrowers with lower outstanding principal balance (OPB), collateral appraisal value (OMV), and higher loan-to-value (LTV) generally tend to exhibit a singular payment spike at a later period. On the contrary, the borrowers with higher OPB, OMV and lower LTV generally demonstrate a faster payment through multiple installments payment.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to Assistant Professor Ekarat Rattagan, Ph.D., whose invaluable guidance, support, and insights have been the cornerstone of this research. Dr. Rattagan's mentorship has not only shaped the direction of this study but has also nurtured my growth and understanding in the academic field.

I am also profoundly thankful to my friends and family. Their unwavering belief in my capabilities, their patience during the trying times of this research, and their encouragement have been instrumental in reaching this milestone in my academic journey.

Additionally, I extend my sincere appreciation to the two private asset management companies that graciously provided the data necessary for this study. Their cooperation and openness have significantly contributed to the depth and authenticity of this research.

To all who have played a part in this academic endeavor, your support has left an indelible mark on both this research and my personal journey in attaining my master’s degree.

Thank you.

Vahrey Sitsuksai

June 2024

TABLE OF CONTENTS

Page

CHAPTER 1
INTRODUCTION

During the Asian Financial Crisis (AFC) in 1997-87, Thailand witnessed a significant increase in non-performing loans (NPLs) ratio, peaking at around 48% in 1999. This surge in NPLs jeopardized the banking system and impeded overall economic development, and the government therefore implemented various policy measures to stabilize the financial system. These include temporary closure and subsequent consolidation of the financial institutions, along with the establishment of an efficient debt restructuring system. The government endorsed the Emergency Decree on Asset Management Company B.E. 2541 (1998), which facilitate the establishment of Asset Management Company (AMC) and paving the way for AMC to acquire defaulted loans and thereby deconsolidate NPLs and non-performing assets (NPA) from the financial system.

Following the acquisition of an NPL portfolio from the selling bank, the AMC engages in a structured settlement process. Settlement can be achieved through cash payment, entry into a troubled debt restructuring (TDR) scheme where borrowers make payments in installments, or the transfer of collateral in the form of assets or equity, which are then restructured and sold in the market. In cases where borrowers refuse to settle, litigation is pursued, and the collateral is sold through auction to third-party buyers or acquired by the AMC using the outstanding loan balance, resulting in its classification as NPA. Once the collateralized asset becomes an NPA, the AMC sells the assets to potential buyers and eventually converts the assets into cash. Figure. 1 provides a visual representation of the AMC's business process and workflow after acquiring a portfolio from the bank.

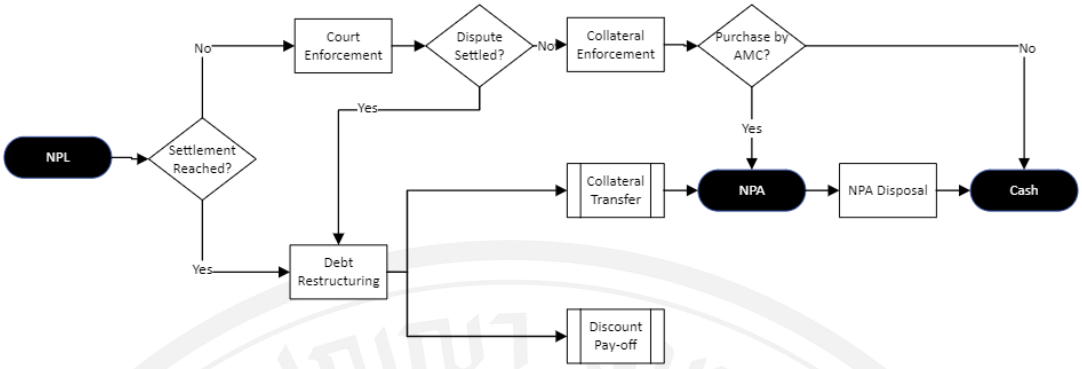


Figure. 1. NPL Management Process in Thailand

The use of AMC in offloading NPLs from the financial system enables financial institutions to grow their loan portfolios, enhance financial stability, and promote economic development. Despite the importance of preserving NPL levels, managing the disposal of these assets can often be fraught with complexity. The dichotomy between the needs of the financial institutions and the AMCs can complicate the resolution process. Financial institutions aim to minimize losses, hence striving to sell their NPLs at the highest possible price. This contrasts with AMCs' objective of purchasing at the lowest possible cost to maximize their profitability. This dynamic often results in a mismatch in valuation pricing (Ciavoliello et al., 2016; Fell, Grodzicki, Martin, & O'Brien, 2017; Pauer & Pichler, 2021) and complicating the negotiation and transaction process between the two parties.

With these complexity, understanding the key determinants of NPL valuation becomes crucial including the recovery rate (Bellotti, Brigo, & Gambetti, 2019; Ye & Bellotti, 2019) and the recovery time (Ciavoliello et al., 2016). The recovery rate is defined as the proportion of money that AMCs successfully collect which comprised the series of cash collection from various period to the OPB upon the loan acquisition. The recovery time is defined as the weighted average number of periods takes the AMC to fully resolve the loan. The time period is range between [0, $+\infty$), the infinity output is for the unresolved cases. These factors together govern the price of NPL portfolios. Accurately estimating these factors may enhance the profitability level of AMCs.

Previous studies have used loan and repayment data to predict the loss given default (LGD) and recovery rate (Bellotti et al., 2019; Cheng & Cirillo, 2018; Ye & Bellotti, 2019), develop models on repayment behavior (Paxton, Graham, & Thraen, 2000), and examine factors that affect loan repayment capability (Bhatt & Tang, 2002). Nevertheless, there is a notable gap in academic research specifically focusing on the recovery time aspect of the loans, which could be attributed to the lack of available data and the complexity of the recovery process.

Despite the extensive studies and efforts directed towards understanding AMC industry and the factors that influence the valuation of NPL, a significant challenge remains: how can AMCs determine and estimate the recovery time of the NPLs? Determining the factors and borrower behaviors influencing recovery time, and discerning the optimal strategies to enhance NPL recovery based on these insights, is a pressing issue yet to be comprehensively addressed in the academic sphere.

This study aims to fill the aforementioned gaps and provide the AMC with a deeper understanding of the factors influencing the recovery patterns of NPLs through cash flow analysis, enabling the development of more effective loan recovery strategies and enhancing the ability to identify borrower behavior patterns and trends associated with varying recovery times.

Therefore, we conduct cluster analysis on the historical collection profiles of borrowers using k-means and agglomerative clustering. During the data transformation, which includes aggregation and normalization, the cash collections will be presented in the form of a time series for each borrower. To facilitate clustering, we will calculate the similarity distance using Dynamic Time Warping, as these are the most common approaches for time series data (Aghabozorgi, Shirkhorshidi, & Wah, 2015). Subsequently, we will apply a random forest classification model to identify the features that explain the clustering results.

1.1 Research Objective

The primary objective of this research is to elucidate the factors influencing the recovery time of NPLs. This understanding is crucial in enabling AMCs to refine their loan underwriting strategies and enhance their loan recovery processes. By identifying and analyzing these key factors, the study aims to contribute to more effective risk management and operational efficiency in NPL management business.

CHAPTER 2
LITERATURE REVIEW

2.1 Cluster Analysis

Clustering is a powerful data mining technique that is used to extract valuable insights from large datasets. With the increasing storage and processing power of modern technology, data is collected in various formats, making it difficult to analyze using traditional methods. Clustering, as an unsupervised learning technique, helps to identify patterns (Lei et al., 2017) and structure in the data, providing a deeper understanding that would be impossible to evaluate using supervised learning algorithms alone.

Clustering is a broad field with a variety of techniques. The methods range from agglomerative and divisive clustering to density-based and model-based techniques. Nevertheless, the basic concept of these approaches remains the same, which is to group similar objects into clusters based on a similarity distance calculation. The goal is to minimize the distance within clusters and maximize the distance between clusters (Ma & Angryk, 2017).

2.1.1 K-Means Clustering

K-means is a widely used unsupervised machine learning technique for data clustering analysis. It partitions the data into a pre-defined number of clusters (k) by grouping similar objects together. The algorithm is centroid-based, meaning that it aims to minimize the sum of distances between each object and its respective centroid. To perform cluster analysis using k-means, one must first select the number of clusters (k) and choose k random points from the data set as initial centroids.

There are several methods to calculate the optimal number of k and silhouette coefficient is a widely used metric for evaluating the quality of clustering algorithms. Silhouette coefficient is calculated using two distance measures: the average dissimilarity between a point and all other points within the same cluster (intra-cluster

distance) and the minimum average dissimilarity between the point and all other points in any other cluster (inter-cluster distance). The result of silhouette coefficient is ranging between -1.0 and 1.0, with -1 indicating that the point should belong to a different cluster and 1 indicating that it is correctly assigned to its cluster.

Subsequent to select the number of k, the algorithm calculates the distance between all other objects and the centroids, assigning each object to the closest centroid. The centroid of each newly formed cluster is then recalculated, and this process is repeated until all objects remain in the cluster, they were previously assigned (Anil K. Jain & Dubes, 1948).

2.1.2 Agglomerative Clustering

Agglomerative clustering is a form of hierarchical clustering which is a widely used technique in data analysis for detecting embedded structures in datasets. It operates on the bottom-up approach by building a hierarchy of clusters by progressively merging smaller clusters into larger ones (Aggarwal & Reddy, 2014). Unlike partitioning clustering methods, agglomerative clustering does not require pre-defined number of clusters. Instead, it treats each data point as a single cluster at the outset and then merges these small clusters into larger and larger clusters, until all points are clustered into a single group or until a specified stopping condition is met.

The process of agglomerative clustering typically begins with the calculation of a similarity or distance matrix, representing the pairwise distances between all points in the dataset. In each iteration, the algorithm merges the pair of clusters that are closest according to the chosen distance. This merging step may follow different linkage criteria, such as single linkage (minimum distance), complete linkage (maximum distance), average linkage, or Ward’s method, which minimizes the variance within each cluster.

The result of agglomerative clustering is often represented as a dendrogram (A.K. Jain, Murty, & Flynn, 1999) that visually represents the merging process and the resulting cluster hierarchy. The approach is especially beneficial for exploratory

data analysis, as it does not impose rigid structures on the data but rather reveals the hierarchical relationships present within the data.

2.1.3 Time-Series Clustering

Time-series clustering is the process of classifying a similar time series into the same cluster based on similarity measures. The technique has been applied to various domains such as astronomy, biology, genetics, climate, psychology, and finance. The method could tackle several real-world problems, such as anomaly detection discovering an unusual and unexpected pattern from the time series data (Leng, Lai, Tan, & Xu, 2009). Pattern discovery is another study frequently observed using the time-series clustering method (Iglesias & Kastner, 2013; Shen & Luo, 2016). A hybrid technique could also lead to prediction and recommendation solutions (Sfetsos & Siriopoulos, 2004).

Time-series is a sequence of continuous nominal values, it comprises many data points. However, when looking at the bigger picture, interesting patterns could be extracted from the connection of those data points. The whole time series could be perceived as a single object (Kumar & Nagabhushan, 2006), and clustering on the entire object is a common method employed by several studies (Aghabozorgi et al., 2015).

There are four components to construct a proper time-series clustering (Aghabozorgi et al., 2015). The first component is the time series representation, which is transforming time-series into the reduced dimensionality vector to a manageable size while preserving an essential character of the original data (Seunghye, 2017). Secondly is the measuring of similarity and distance. The two most common approaches to calculating the similarity in time series are Euclidean distance and DTW.

Euclidean distance is a classical method for calculating the distance under the k-means clustering problem. The approach is surprisingly competitive when there is low dimensionality data, the data has an equal length of time, and the objective of clustering is to measure similarity in time (Aghabozorgi et al., 2015). If there are two

time series s and u consisting of T samples each $s = (s_1, s_2, s_3, \dots, s_t) \in S^T$ and $u = (u_1, u_2, u_3, \dots, u_t) \in U^T$, then the squared Euclidean distance between two time series s and u is given by:

$$d_E(s,u) = \sqrt{\sum_{t=1}^{T} (s_t - u_t)^2}$$

On the contrary, DTW is more appropriate when the timing factor is less critical and pattern extraction is the objective of the clustering exercise. DTW can generate an optimal global alignment between two time series, allowing the similar shape to match even if there is a temporal distortion between time series (Cassisi, Montalto, Aliotta, Cannata, & Pulvirenti, 2012). Below is the formula to calculate the distance under DTW approach:

$$d_{DTW}(s,u) = min_{\pi} \sqrt{\sum_{(i,j)\in\pi} d(s_i, u_j)^2}$$

where $\pi = [\pi_0, \dots \pi_K]$ is a path that satisfies the following properties:

A list of index pairs $\pi_k = [i_k, j_k]$ with $0 \le i_k < n$ and $0 \le j_k < m$

- $\pi_0 = (0,0)$ and $\pi_{K-1} = [n-1, m-1]$
- For all $k > 0, \pi_k = (i_k, j_k)$ is related to $\pi_{k-1} = (i_{k-1}, j_{k-1})$ as follows:
    - $i_{k-1} \le i_k \le i_{k-1} + 1$
    - $j_{k-1} \le j_k \le j_{k-1} + 1$
- n and m is the length of the first and second time series, respectively

The third component is the cluster prototypes or also known as cluster representatives, and the prototype will be the factor determining the quality of the cluster when performing the analysis. There are generally three approaches to defining the cluster representative: the mean of the sequence set, the medoid of the sequence set, and the local search prototype. The last component is the clustering algorithm, which is similar to the general clustering algorithms, for example, hierarchical clustering, partitioning clustering, grid-based, model-based, density-based, and multi-step clustering.

Figure. 2 presents the visualization of the distance calculation between Euclidean distance (left) and DTW distance (right).

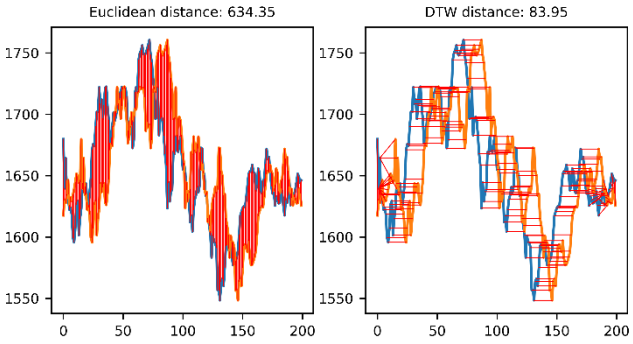


Figure. 2. Illustration of distance calculation applying different approaches

2.2 Random Forest Classification Model

Employing a classification model to ascertain the key features that contribute to explaining the clustering results (Brieman, 2001) holds considerable value in the context of this research. There are several traditional classification models including logistic regression, decision tree, k-nearest neighbor, neural network and random forest. As a powerful and widely adopted machine learning algorithm, the random forest model has demonstrated its effectiveness in determining the most significant features influencing the result. Nonetheless, random forest operates by constructing a multitude of decision trees during the training phase and generating a classification output based on the mode of the classes produced by individual trees. Moreover, the random forest classifier offers the built-in feature importance ranking, which can be beneficial in understanding the contribution of each variable to the classification process (Liaw & Wiener, 2002).

CHAPTER 3
RESEARCH DESIGN AND METHODOLOGY

3.1 Research Methodology

It is important to note that the data used in this study only includes borrowers who have fully resolved and therefore the value of recovery time result should be in a range of 0 to less than infinity, allowing for normalization using total collections. From this point forward, the aggregated and normalized cash collection will be referred to as ‘cash flow’. This term will be used consistently throughout the remainder of the paper to represent the combined cash collection values adjusted for normalization.

The cash flow data will serve as the primary input for conducting time-series clustering, utilizing both k-means clustering and agglomerative clustering approaches. To calculate the similarity distance between different data points, we applied both the Euclidean distance and Dynamic Time Warping techniques, which are two of the most common methods for working with time-series data (Aghabozorgi et al., 2015). After obtaining the clusters of cash flow, these results will be used as input data in a random forest classification model to further explore the features that contribute to different cluster results. The random forest model combines multiple decision trees, which reduces the risk of overfitting. Additionally, its ability to provide insights into feature importance is critical for understanding which variables are driving the model's predictions.

Figure. 3 delineates the methodology adopted in this study, which collectively contribute to the robust and comprehensive exploration of the research subject at hand. The first step involves Input 1, where the initial time series data is collected. This data undergoes Process 1, where time series clustering is applied, resulting in Output 1. Subsequently, additional data which is comprises of the characteristics of the borrowers, denoted as Input 2, is incorporated. Both Output 1 and Input 2 serve as inputs for Process 2, where a Random Forest Classification Model is applied to

extract significant features, leading to Output 2. Finally, the results from Output 2 are subjected to thorough analysis to identify the features that could explain the clustering result.

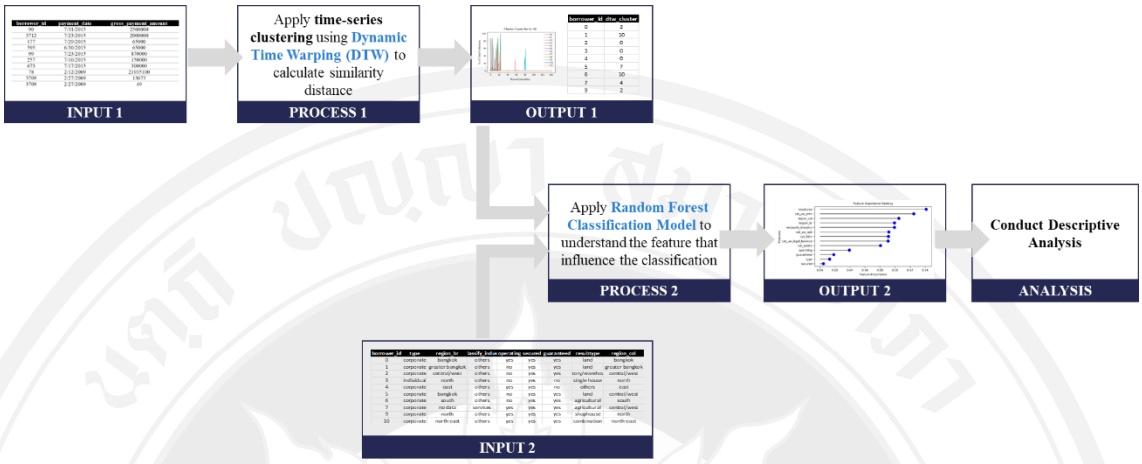


Figure. 3. Flow diagram of the proposed research approach

3.2 Data Set

The dataset used in this study is a real-world dataset obtained from two private asset management companies in Thailand. The dataset comprises underwriting and post-acquisition statistics from 2008 to 2020, with a total of 2,839 borrowers serving as input for the time-series clustering. Additionally, there are 13 features linked to the borrower id which will be analyzed to determine the characteristics of the borrower.

3.2.1 Raw Data for Time-Series Clustering

This dataset is the actual cash collection by period which is the amount of cash that AMC could collect from the borrowers or liquidate the collateralized asset related to the borrower within each period. Table 1. is example of raw data before preprocessing.

Table 1. Raw Cash Collection Data

borrower_id	payment_date	gross_payment_amount
90	7/31/2015	2500000
3712	7/23/2015	2000000
177	7/29/2015	65000
595	6/30/2015	65000
99	7/23/2015	870000
257	7/10/2015	150000
673	7/17/2015	300000
78	2/12/2009	21935100
3709	2/27/2009	13673
3709	2/27/2009	49

Table 1. provides a detailed overview of borrower payment behavior. An illustration of this can be found in row 1, where borrower 90's transaction with the AMC is presented. This borrower made a substantial payment of 2.5 million baht on 31 July 2015. It is noteworthy to mention that a borrower's payments could be arranged in diverse ways. For instance, they might remit one substantial sum or opt for several smaller transactions. This is demonstrated in the final two rows of the table, highlighting the payment activity of borrower 3709 on 27 February 2009. Despite both transactions occurring on the same day, they are listed separately, with amounts of 13,673 baht and 49 baht respectively.

3.2.2 Raw Data for Cluster Analysis

Table 2. are set of information will comprise of socio-demographic and loan file and collateral information of the borrower.

Table 2. Selected Features for Classification and Data Analysis

Category	Features	Example of Data
Borrower Information	1. Borrower Type	• Corporate • Individual
	2. Industry	• Real Estate • Construction • Finance

Category	Features	Example of Data
		• Manufacturing
	3. Guaranteed	• Yes / No
	4. Operating	• Yes / No
	5. Secured	• Yes / No
	6. Residing Location – Region	• Bangkok • Greater Bangkok
Collateral Information	7. Collateral Sub Type	• Condominium • Single House • Detached House
	8. Region	• Bangkok • Greater Bangkok
Loan File	9. Legal Balance	• 8,000,000
	10. OPB	• 7,000,000
	11. Appraisal Value	• 8,500,000
	12. OPB-to-Value	• 82.3%
	13. LB-to-Value	• 87.5%

The main socio-demographic features included in this database were the borrower type (individual vs. corporate), operating industry (construction, finance, services, trading or etc). Loan file information refers to the outstanding loan value at the default event and its collateralized assets such as OPB, legal balance (LB), and loan-to-value (LTV) of the loans. The collateral information includes type of collaterals (single house, detached house, townhouse, shop house, land, or warehouse), collateral location (region, province, district), size of the assets, appraisal value of the assets.

3.3 Data Preprocessing Procedure

3.3.1 Preprocessing Input Data

Data preprocessing were employed to clean, transform, and standardize the dataset, ensuring its suitability for accurate and reliable analysis. The following methodologies were implemented to prepare the data for subsequent clustering and classification models.

3.3.2 Data for Time-Series Clustering

Aggregation: Transforming the raw data into a series of collection patterns by aggregating the daily collection from Table 1. into a monthly collection for each borrower to reduce computation time and with the time scale that is too small, global trends may be difficult to discover (Stetco, Zeng, & Kaene, 2013). It is important to note that each borrower may have a different payment pattern, some might make a payment in lumpsum single payment, or making payment in installment. Hence, by doing the aggregation, the information for each individual borrower is aggregated into a single row.

Normalization: The next step involved in the analysis was to normalize the data, which is a process of standardizing the time series of each borrower. This step helps to address the issue of major deviations in absolute value that may exist within the data. To achieve this, we used a formula to normalize the collection of each period, which allowed us to put all the data on a common scale for easier comparison and analysis.

$$normalized\_C_t^i = \frac{C_t^i}{Total\ collection\ from\ Borrower\ i}$$

Whereby: C_t^i *is a collection of borrower i at time t*

By normalizing the time series data of each borrower, we were able to eliminate any variations in absolute value that could have had an impact on our analysis. This led to a transformation of the data from Table 1 into an illustrative format that could be more easily analyzed in Figure. 4 below.

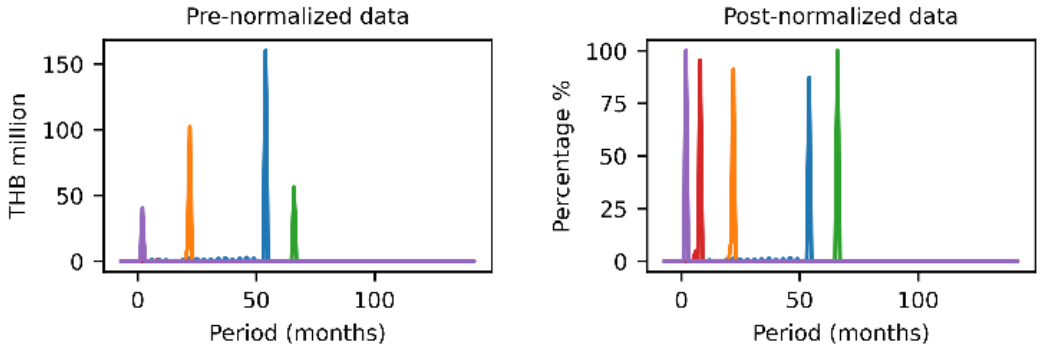


Figure. 4. Pre and Post-Normalized Time-Series Data

The post-normalized data was represented on a right graph, with the y-axis indicating 100% of the collection of that particular borrower and the x-axis representing the collection timing. The graph has set the starting point, month zero, to be the same as the month in which the loan was acquired. Since the data was collected from 2008 to 2020, the x-axis has been limited to 144 months, equivalent to 12 years.

3.3.3 Correlation Coefficient of Independent Features

Before deploying the classification model, a correlation analysis was conducted to identify features that are either redundant or irrelevant. This step helps eliminate variables (i.e. ‘region_br’, ‘secured’, ‘uw_legal_balance’, and ‘lbltv’) that could unnecessarily increase the variance of the coefficient estimates, thereby leading to potential instability in the model's predictions. Figure. 5 is the correlation analysis between original features.

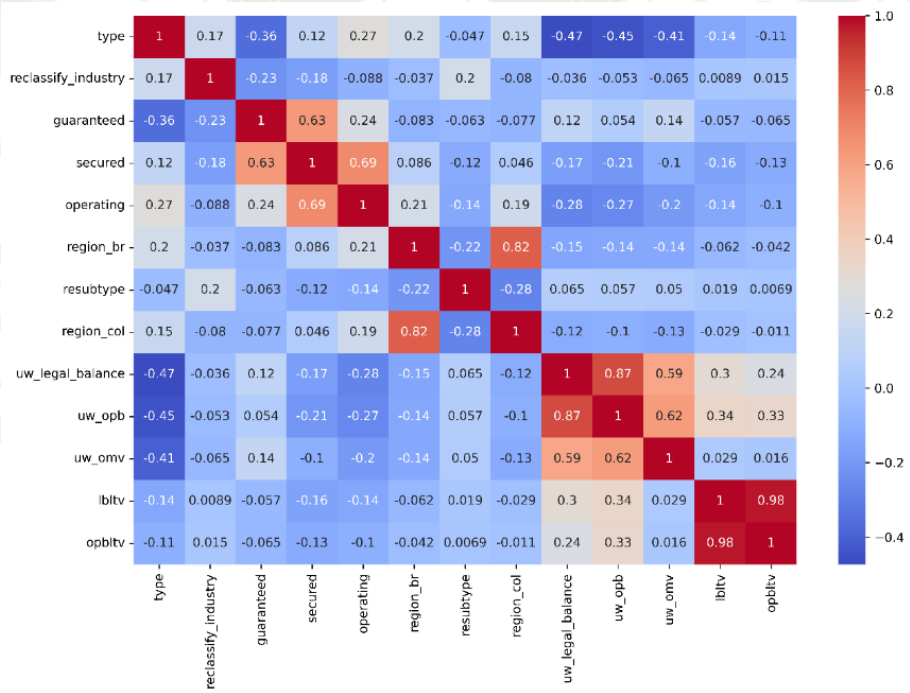


Figure. 5. Correlation Analysis

The findings indicate that the borrower's residing location ('region_br') and the collateral's location ('region_col') exhibit a substantial linear relationship, as evidenced by a correlation coefficient of 0.82. Furthermore, the 'secured' feature is positively correlated with both 'operating' and 'guaranteed' features, with respective correlation coefficients of 0.69 and 0.63. These relationships highlight the potential interconnectedness of geographical factors and loan security characteristics in the dataset.

In term of loan file data, the correlation result reveals a high correlation between the underwriting legal balance (‘uw_legal_balance’), underwriting outstanding principal balance (‘uw_opb') and underwriting appraisal value ('uw_omv’), reflected by a correlation coefficient of over 0.59. Moreover, the loan-to-value ratio (‘ltv’) base of legal balance and OPB displays a correlation coefficient of 0.98. Given these findings, it may be prudent to consider the exclusion of the ‘region_br’, ‘secured’, ‘uw_legal_balance’, and ‘lbltv’ features from subsequent modeling to avoid redundancy, as they demonstrate substantial correlation with other variables within the dataset.

3.4 Time-Series Clustering and Classification Model Application

As we sought to segment the time series data, the historical cash collection, into distinct clusters, we will apply two distinct time-series clustering approaches: k-means clustering and the agglomerative clustering model. This analysis aims to provide valuable insights into the repayment behavior of the borrowers in different clusters.

In our initial attempts, we applied the Euclidean distance for distance calculation in k-means clustering; however, the results were not up to the mark, registering a negative silhouette score that ranged between -0.290 to -0.390. This was not the outcome we were aiming for, as it indicated that the distance within the clusters was greater than the distance between different clusters. On the other hand, the DTW method yielded more satisfactory results, with scores spanning from 0.536 to 0.691. Hence, we utilized the DTW method to measure the similarity for time-

series clustering on the post-normalized data on both clustering approach which we will named it as DTW-based k-means clustering and DTW-based agglomerative clustering. This clustering analysis sets the foundation for further exploration of the factors influencing the recovery patterns.

After obtaining the clustering result, random forest classification model will later be applied to identify the features that influences the clustering results. This would further enable targeted analysis of select features to understand the distinct characteristics of each group within the cluster.

3.4.1 DTW-based K-means Clustering

While the k-means approach necessitates a pre-determined number of clusters, the silhouette coefficient was utilized to pinpoint the optimal number of clusters for further scrutiny. In order to expedite the computation of the silhouette coefficients, we randomly chose 40% of the complete cash flow data to perform the analysis. The output of the computation is visualized in Figure. 6 below.

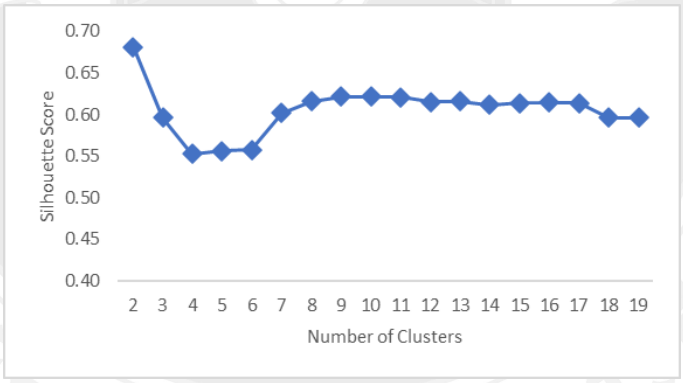


Figure. 6. Silhouette score from DTW-based k-means clustering

According to the silhouette coefficient analysis, the top three optimal number of clusters (k) appears to be 2, 9, and 10 with the silhouette coefficient score of 0.6812, 0.6216, and 0.6216, respectively. The result demonstrate that these number of cluster has the highest degree of alignment of members within their respective clusters. Furthermore, considering the similarity in the results for k=9 and k=10, the

analysis on clustering therefore will concentrate exclusively on k=2 and k=9. This decision is based on the objective of achieving a more focused and efficient examination of the clustering outcomes.

There are two primary outcomes have emerged from the cluster result, first is the identification of each group's center cluster, calculate using the barycenter averaging method (tslearn, 2017). The center cluster will be use as the representation of the collection timing of different clusters. The second outcome involves the allocation of individual borrowers to these clusters, providing a basis for subsequent classification modeling by leveraging this clustering information.

Table 3. Clustering Result for k=2 and k=9

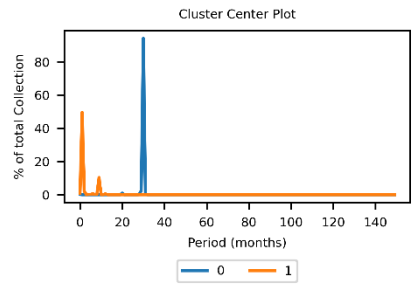
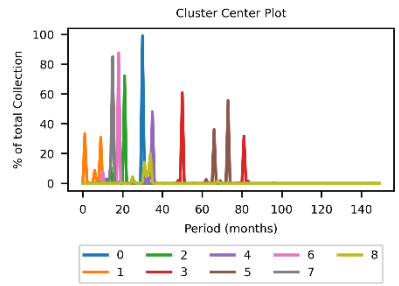
k	Cluster Center	Recovery Months			
2		**group**	**member**	**month**	**center summation**
		0	2,355	30.21	97.90
		1	484	2.25	64.36
9		**group**	**member**	**month**	**center summation**
		0	1,485	30.91	99.80
		1	59	5.07	80.70
		2	200	19.28	93.17
		3	66	62.32	100.54
		4	113	22.10	62.27
		5	100	70.13	98.84
		6	451	17.95	98.60
		7	320	15.65	99.03
		8	45	18.94	55.70

Table 3. above illustrates the clustering result for k=2 and k=9, whereby, it is observe that majority of the members are grouped into cluster 0 based on their recovery patterns, which have a weighted average recovery time of 30.2 and 30.9 months, or around 2.5 years, respectively. Considering that the study’s input data

comprises exclusively fully recovered loans, it is expected that the sum of the cluster centers post-normalization should approximate 100. Nevertheless, the observed results reveal that certain cluster groups have cumulative center significantly below this threshold, which could potentially leading to deviations in the analysis of absolute recovery times. However, since the primary focus of this study is on analyzing the recovery patterns of cluster groups rather than the absolute recovery times, the decision has been made to retain these groups in the analysis and proceed with the result of entire dataset.

For demonstration purposes, the provided cluster center plot for k=2 elucidates distinct payment behaviors within the two clusters. For Group 0, denoted by the blue line, there is a prominent peak at the 30-month interval. This sharp spike suggests a tendency among members of this group to make a significant one-time payment at this particular juncture. Conversely, group 1, represented by the orange line, exhibits two early spikes, which are less pronounced than the spike observed in Group 0. This pattern may suggest that borrowers in Group 1 are more likely to start making payments earlier compared to Group 0. However, the payments appear to be split over multiple transactions, as evidenced by the multiple spikes.

In the k=9 clustering configuration, the data reveal a discernible variability, suggesting that this finer granularity effectively captures diverse borrower behaviors. Notably, Group 0 in the k=9 cluster exhibits a recovery pattern akin to Group 0 in the k=2 cluster, characterized by a pronounced payment spike around the 30-month mark. The distribution of average recovery times across the k=9 clusters demonstrates a broad spectrum of loan repayment durations, which enhances the understanding of the varied payment timelines adhered to by different borrower segments. Such insights could be pivotal for developing targeted loan recovery strategies and refining risk assessment models. Nonetheless, it is important to reemphasize that with respect to the silhouette coefficient, the result for k=9 is marginally lower than that for k=2, suggesting that the cluster cohesion and separation may not be as pronounced which could potentially reflect a more complex and overlapping borrower behavior structure.

As a result, contrasting payment patterns from the cluster analysis offer significant insights into the financial behavior and payment strategies of the borrowers within each cluster. Such information can serve as a critical input for a more in-depth analysis of borrower characteristics across the different groups. Understanding these nuances enables the development of tailored approaches to loan management and recovery, potentially leading to more effective financial practices and policies.

3.4.2 DTW-based Agglomerative Clustering

In the context of hierarchical clustering models, there is no readily available function for computing the DTW similarity distance. Therefore, the calculation of these distances is performed in advance by applying ‘dtw metrics’ function from tslearn library in python.

Subsequently, the DTW distance calculation was used as an input into a hierarchical clustering model, employing the Ward linkage criterion. The rationale behind applying this methodology lies in its capacity to minimize the within-cluster variance. This approach ensures that the selected pair of clusters to be merged at each step are chosen in a manner that minimizes the increase in total within-cluster variance post-merger, which leads to a well-defined cluster. The result of hierarchical clustering was presented below in Figure. 7 in a form of Dendogram.

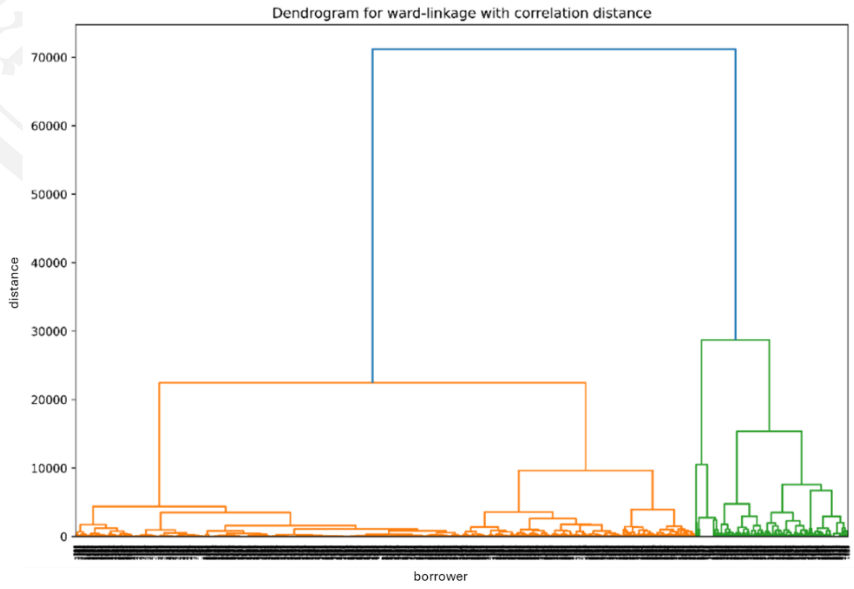


Figure. 7. Result of DTW-based Agglomerative Clustering

The outcomes of the hierarchical clustering analysis suggest that dividing the dataset into two distinct groups is optimal. This finding aligns closely with the results obtained from the k-means clustering analysis, where the silhouette score for k=2 emerged as the most robust, further reinforcing the suitability of partitioning the data into two clusters. Moreover, cophenetic correlation coefficient calculations have been applied to verify the correlation between the datapoints in the dendrogram, with the model generating a correlation coefficient result of 0.8294, indicating that the clustering algorithm was well preserve the original data structure.

In agglomerative clustering, unlike k-means clustering which identifies a centroid as the representative of each group, there is no direct provision for determining the cluster center. Therefore, to represent each cluster, the medoid of the cluster will be selected. The following is a summary of the results obtained from agglomerative clustering with k=2.

Table 4. Agglomerative Clustering Result for k=2

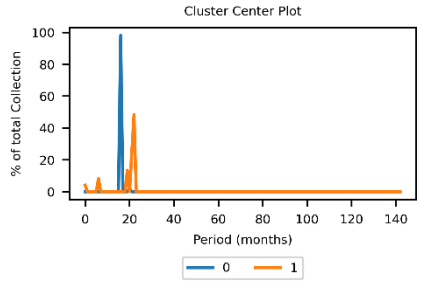
k	Cluster Medoids	Recovery Months			
		group	member	month	center summation
2		0	2,278	16.07	100.00
		1	561	19.10	98.25

Table 4. illustrates the clustering results for k=2, presenting the cluster medoids for each group. It is observed that the majority of members are allocated to Cluster 0 based on their recovery patterns, with the cluster medoids indicating average recovery times of 16.07 and 19.10 months, respectively. The medoids reveal distinct

payment behaviors between the two clusters. For the medoid of Group 0, represented by the blue line, a significant one-time payment is observed at month 16, indicating that this borrower opts to repay a lump sum amount 16 months after loan acquisition. In contrast, the medoid of Group 1, depicted by the orange line, displays multiple spikes, with a smaller spike in the early days and a larger spike later on. This pattern suggests that the borrower attempts to make installment payments early in the period and aims to settle the transaction later.

Although the medoids identified through agglomerative clustering indicate a shorter average recovery time, the payment patterns of the borrowers closely align with the results from k-means clustering. The majority of members in Group 0 exhibit one-off payments, while the second cluster demonstrates a pattern of multiple payments. Additionally, the distribution of members across each cluster is remarkably similar between both clustering methods.

3.4.3 Classification Model

Given the similarity in outcomes between k-means and agglomerative clustering, we have decided to focus exclusively on the results obtained from k-means clustering for the subsequent stages of our analysis. This decision is motivated by the objective to employ a classification model that elucidates the characteristics distinguishing borrowers across different groups. Concentrating on the k-means clustering results will streamline our analysis, enabling a more directed exploration of borrower behaviors within the identified clusters.

The clustering results present an imbalanced data issue, which may lead to biased model performance where the classifier favors the majority class while neglecting the minority class. Employing upsampling techniques enables the creation of a more balanced dataset, which, in turn, improves model performance and ensures that the classifier effectively captures the underlying patterns of both majority and minority classes. In this research, the original sample size has been upsampled for all clusters, excluding group zero, to make the sample sizes of the other clusters almost match that of group zero, thereby addressing the class imbalance issue.

To ensure fair representation across all classes, the researcher trained the classification model using upsampled data to enhance the model's performance, particularly for underrepresented classes. Subsequently, the model was evaluated on the original data (non-upsampled data) to obtain a realistic estimate of its performance, providing an accurate representation of how the model would perform in real-world scenarios where class imbalances might still be present.

To evaluate the performance of the classification model, the log-loss metric was employed for each identified cluster. A lower log-loss score signifies the model's enhanced capability to accurately assign the correct cluster membership to individual observations. Figure. 8 below presents the log-loss scores obtained from applying a random forest classification model to the results of each clustering configuration.

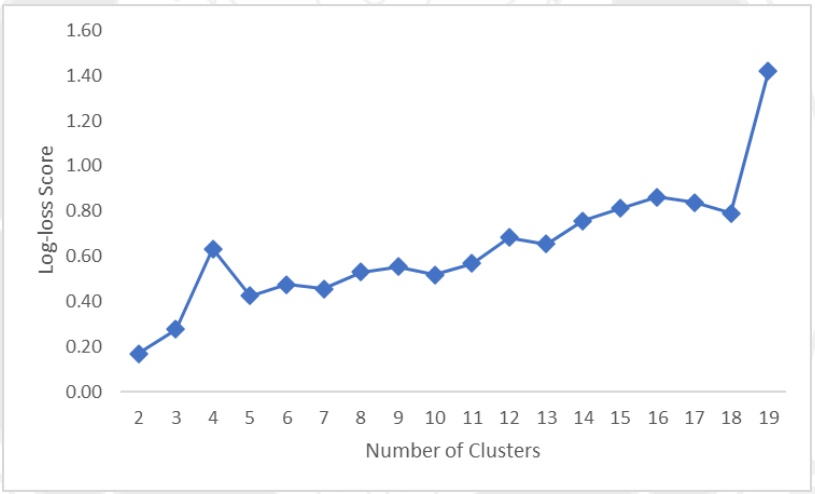


Figure. 8. Log-loss score based on k-means clustering results

The classification model's performance, evaluated using the log-loss metric across varying numbers of clusters, demonstrates that the most accurate predictions are achieved with a bisection of the dataset (k=2), where the log-loss score is the lowest at 0.17. This indicates a high confidence in the model's probabilistic predictions. As the number of clusters increases, a general uptrend in log-loss scores is observed, suggesting diminishing predictive accuracy. Notably, there is a pronounced escalation in log-loss scores beginning at k=12, culminating in the

highest score of 1.42 at k=19, which could be indicative of overfitting. The fluctuating scores between k=6 and k=11 imply that while the model can somewhat adapt to additional complexity, there exists an optimal threshold for the number of clusters beyond which the accuracy of predictions degrades.

The secondary outcome derived from the classification model pertains to the analysis of feature importance. The significance of this analysis lies in the interpretation of the scores, where higher values denote a greater influence on the classification model's predictive capabilities. Contrary to the variability observed in the log-loss scores, the evaluation of feature importance across the clusters reveals a striking uniformity. As presented in Table 5. the features across almost every cluster exhibit similar levels of importance, with underwriting OPB (‘uw_opb’), loan-to-value based on OPB (‘opbltv’), and underwriting appraisal value (‘uw_omv’) demonstrating the most significant impact to do model prediction. The three features altogether contribute around 0.71 – 0.73 of the model prediction result. This consistency in feature significance underscores the pivotal role these elements play in the model's ability to discern between different classes, suggesting their critical influence in the classification process regardless of the separation of number of clusters and log-loss outcomes.

Table 5. Feature Importance Ranking

	Number of Clusters																	
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
uw_opb	0.26	0.26	0.25	0.26	0.25	0.25	0.25	0.25	0.25	0.24	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
opbltv	0.24	0.23	0.23	0.24	0.23	0.24	0.23	0.23	0.23	0.24	0.23	0.23	0.23	0.23	0.23	0.24	0.23	0.23
uw_omv	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.22	0.23	0.23	0.23	0.23	0.23	0.23	0.23
region_col	0.08	0.09	0.09	0.09	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.09	0.09	0.09	0.09	0.09	0.10
reclassify_industry	0.06	0.07	0.06	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.08	0.07	0.07	0.07
resubtype	0.05	0.05	0.06	0.05	0.05	0.05	0.06	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
operating	0.04	0.04	0.04	0.04	0.04	0.04	0.03	0.04	0.03	0.03	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
guaranteed	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
type	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

CHAPTER 4
RESULTS AND DISCUSSION

In this section, we focus on the findings from a classification model based on the results of k-means clustering. Specifically, our analysis centers on the results obtained from using two clusters (k=2). This choice is grounded in the fact that k=2 not only demonstrated the highest silhouette score in the k-means clustering, but also yielded the lowest log-loss score in the classification model. The low log-loss score suggests that the model is particularly effective in making accurate predictions within this group configuration. This dual criterion of high silhouette and low log-loss scores underpins our decision to concentrate on the k=2 cluster result for a more detailed examination.

In the classification results, it is evident that uw_opb (underwriting outstanding principal balance), opbltv (loan-to-value ratio), and uw_omv (underwriting original market value) emerge as the most significant contributors to the model's predictive capacity. To further explore these findings, we have provided a boxplot below, excluding the outliers value, which visually delineates the distribution range across the different clusters.

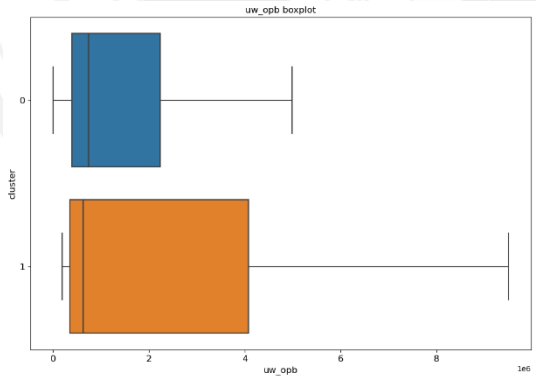


Figure. 9. Boxplot on uw_opb

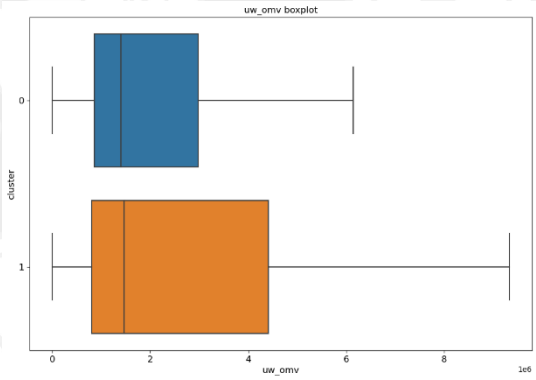


Figure. 10. Boxplot on uw_omv

The boxplot analysis of 'uw_opb' as shown in Figure. 9, reveals distinct financial characteristics among the grouped members across the two clusters. Cluster 0, comprising 2,321 individuals, displays a mean loan balance of approximately 5.4

million. Conversely, Cluster 1, though smaller with 446 members, is characterized by a higher average loan balance of about 11.3 million. Nonetheless, the boxplot analysis shows that Cluster 1 has a wider interquartile range compared to Cluster 0, indicating a greater spread among the central 50% of loans and reflecting a more pronounced heterogeneity within this cluster.

When considering ‘uw_omv’ shown in Figure. 10, a similar trend emerges. Cluster 0 shows an average market value of roughly 3.97 million, whereas Cluster 1's average market value is higher, at approximately 5.69 million, underscoring the higher asset value in Cluster 1. However, despite the higher mean value in Cluster 1, both clusters display significant variability, as indicated by their standard deviations. This variability is reflective of a wide range of asset values within each cluster. The broader interquartile range in Cluster 1 for both 'uw_opb' and 'uw_omv' underscores the diversity in financial and asset valuation within the cluster, illustrating the complex financial landscapes these borrowers navigate.

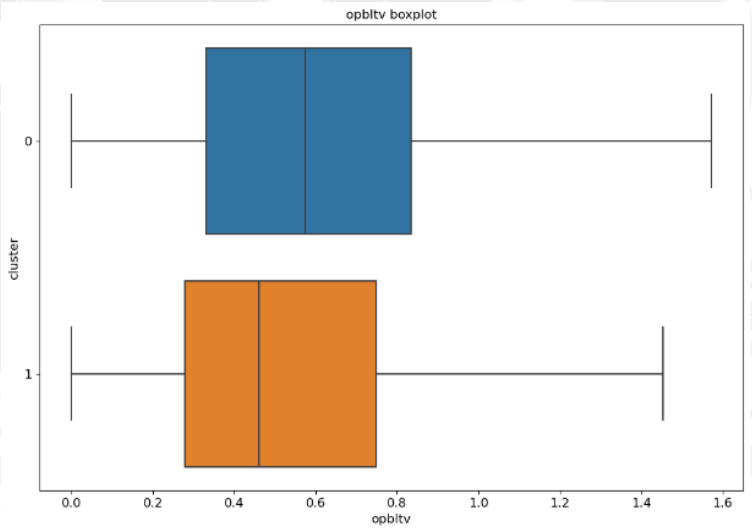


Figure. 11. Boxplot on Loan-to-Value Ratio

The analysis of the loan-to-value ratio presented in Figure. 11 has clarified the disparities between the clusters, revealing that Cluster 1 generally exhibits a lower loan-to-value ratio compared to Cluster 0, which displays a ratio that is approximately 10% higher across all measures. This finding aligns with business insight that

borrowers with lower LTV ratios are more inclined to actively engage in repaying their loans to the AMC as the value of their assets significantly exceeds their loan amount. This situation often results in payments being made in multiple tranches, as reflected in the center cluster of multiple spikes of payment, reflecting the borrowers' commitment to fulfilling their obligations. In contrast, the singular payment spike observed in Cluster 0 may suggest that these borrowers opt for one-off payments, possibly through deed transfer or asset divestment, indicating a different approach to loan settlement.

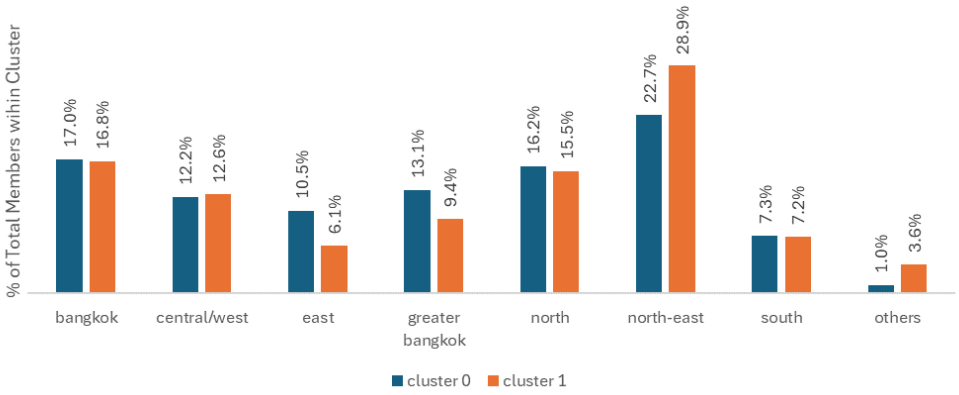


Figure. 12. Distribution of the members by region of the collateral

The feature 'region_col,' denoting the region of the collateralized assets, emerges as the fourth most influential in the analysis. The analysis of the feature as shown in Figure. 12, reveals distinct geographical distributions among the clusters. Regions such as Bangkok and Central/West exhibit a relatively balanced distribution of members across both clusters. In contrast, regions like the East, Greater Bangkok, and the Northeast, as well as the category labeled "Others," demonstrate more marked disparities in membership between Cluster 0 and Cluster 1.

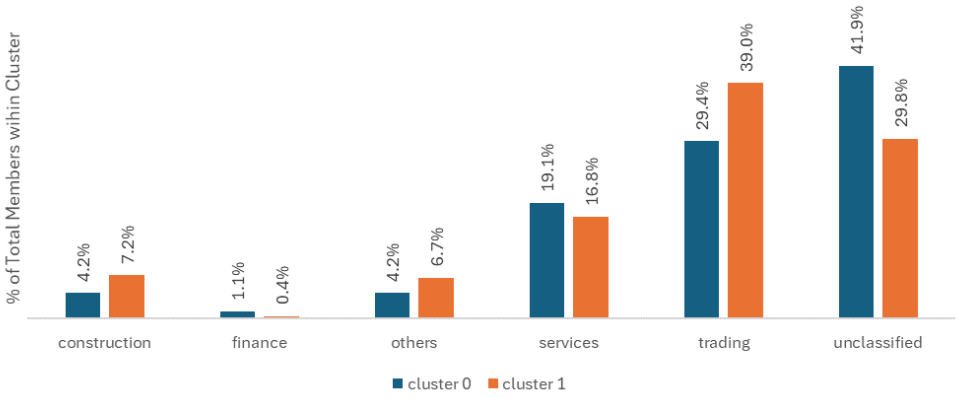


Figure. 13. Distribution of the members by operating industry

The attribute 'reclassify_industry,' as shown in Figure. 13 presents the classification of the industry that the borrowers operates in at the time of loan acquisition, presents noteworthy trends despite a considerable proportion of data being categorized as 'unclassified' due to missing information. A prominent trend emerges within the 'trading' category, where 39.0% of the members in Cluster 1 are associated with this sector. This observation suggests that borrowers engaged in the trading industry may have a higher propensity to be grouped into Cluster 1, indicating a potential industry-related pattern in the clustering results.

With the findings presented in this study, the AMC may be able to refine its underwriting strategy. Such refinement might enable the AMC to forecast recovery timelines with greater precision, leading to improved capital allocation, better resource management, and the setting of more realistic expectations for stakeholders.

For instance, if an AMC observes an NPL portfolio with predominant features that align with faster recovery patterns based on this research, it might decide to allocate more resources or offer a premium during acquisition. Conversely, portfolios that exhibit characteristics linked with slower recovery times might be approached with more caution or subject to more rigorous underwriting to manage the associated risks.

Furthermore, the AMC may tailor loan recovery strategies for borrowers of varied profiles, aiming to enhance collection efficiency.

CHAPTER 5
CONCLUSION

This study aims to develop an understanding of recovery time of NPL based on collection patterns. The data involved in this research includes 2,839 loans. The analysis required preprocessing, normalization, and transformation of the raw data, followed by time-series clustering using DTW for similarity distance calculation. Subsequently, the results were passed through a random forest classification model to identify the features that significantly impact the clustering results.

This research encompasses a two-stage analysis, first the DTW-based clustering result and secondly the classification result. Under the clustering result, it has been demonstrated that dividing the members into two distinct groups yields the most optimal outcomes, regardless of whether k-means clustering, or agglomerative clustering is applied. Specifically, within the k-means clustering framework, the highest silhouette score was observed at k=2, indicating an optimal grouping of members into two distinct categories. Similarly, agglomerative clustering, employing a bottom-up approach as illustrated by the dendrogram, also suggests that members are most effectively differentiated into two groups. This is further supported by a cophenetic correlation coefficient of 0.8294, underscoring the effectiveness of this bifurcation in preserving the original data structure.

Furthermore, the application of the classification model suggests increased predictive accuracy when the potential outcomes are fewer, as is the case with k=2, which is indicated by the lowest log-loss score. However, regardless of the number of clusters or variations in the log-loss score, the features contributing to the model's predictability remain consistent across all models. The feature importance ranking reveals that the underwriting outstanding principal balance, loan-to-value ratio, and underwriting appraisal value of the collateral assets are the predominant features, collectively accounting for over 70% of the influence on the cluster profiles.

The study delineates clear distinctions in payment patterns based on these financial indicators. Borrowers characterized by a lower OPB and appraisal value,

alongside a higher LTV, tend to exhibit a singular payment spike as observed in Cluster 0. This behavior suggests a tendency towards lump-sum payments within a defined period after loan acquisition.

Conversely, borrowers with a slightly higher OPB and appraisal value, coupled with a lower LTV, generally demonstrate quicker repayment through multiple installments. This pattern implies a more proactive approach to loan settlement and could reflect better financial stability or strategic financial management by these borrowers.

It is pertinent to note that 66.3% of the dataset utilized in this research pertains to portfolios acquired in 2017. As a result, the findings might not adequately encapsulate the entire NPL management cycle, which typically spans over five to ten years. Moreover, this study does not account for external and macroeconomic factors that may influence the recovery patterns of borrowers.

Future research could benefit from incorporating these external factors to examine the elements that influence the NPL recovery time, as well as conducting a focused analysis on portfolios that have completed the entire NPL cycle for more insightful results.

BIBLIOGRAPHY

Aggarwal, C. C., & Reddy, C. K. (2014). Data Clustering Algorithms and Applications. In *Data Clustering Algorithms and Applications* (pp. 100).

Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015). Time-series clustering - A decade review. *Information Systems, 53*, 16-38.

Bellotti, A., Brigo, D., & Gambetti, P. (2019). Forecasting Recovery Rates on Non-Performing Loans with Machine Learning. *Credit Scoring and Credit Control Conference XVI*.

Bhatt, N., & Tang, S. Y. (2002). Determinants of Repayment in Microcredit: Evidence from Programs in the United States. *International Journal of Urban and Regional Research, 26*(2), 360 - 376.

Brieman, L. (2001). Random Forests. *Machine Learning, 45*, 5–32.

Cassisi, C., Montalto, P., Aliotta, M. A., Cannata, A., & Pulvirenti, A. (2012). Similarity measures and dimensionality reduction technique for time series data mining. In *Advances in data mining knowledge discovery and applications* (pp. 70-96): InTech.

Cheng, D., & Cirillo, P. (2018). A Reinforced Urn Process Modeling of Recovery Rates and Recovery Times. *Journal of Banking and Finance, Forthcoming*.

Ciavoliello, L. G., Ciocchetta, F., Conti, F. M., Guida, I., Rendina, A., & Santini, G. (2016). *What's the value of NPLs.* Retrieved from

Fell, J., Grodzicki, M., Martin, R., & O'Brien, E. (2017). A Role for Systemic Asset Management Companies in Solving Europe's Non-Performing Loan Problems. *European Economy Banks*.

Iglesias, F., & Kastner, W. (2013). Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns. *Energies, 6(2)*(Hybrid Advanced Techniques for Forecasting in Energy Sector), 579-597.

Jain, A. K., & Dubes, R. C. (1948). Algorithms for Clustering Data. In (pp. 96-97). New Jersey: Prentice-Hall.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data Clustering: A Review. *ACM Computing Surveys, 31*, 60.

Kumar, R. P., & Nagabhushan, P. (2006). *Time Series as a Point - A Novel Approach for Time Series Cluster Visualization.* Paper presented at the Conference on Data Mining.

Lei, Y., Bezdek, J. C., Romano, S., Nguyen, X. V., Chan, J., & Bailey, J. (2017). Ground truth bias in external cluster validity indices. *Pattern recognition, 65*, 58-70.

Leng, M., Lai, X., Tan, G., & Xu, X. (2009). *Time Series Representation for Anomaly Detection.* Paper presented at the IEEE International Conference on Computer Science and Information Technology, Beijing, China.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News, 2*(3), 18-22.

Ma, R., & Angryk, R. A. (2017). *Distance and Density Clustering for Time Series Data.* Paper presented at the IEEE International Conference on Data Mining Workshops.

Pauer, F., & Pichler, S. (2021). Sell or Hold? On the Value of Non-Performing Loans and Mandatory Write-Off Rules. *SSRN*.

Paxton, J., Graham, D., & Thraen, C. (2000). Modeling Group Loan Repayment

Behavior: New Insights from Burkina Faso. *Economic Development and Cultural Change, 48*.

Seunghye, W. J. (2017). Data representation for time series data mining: time domain approaches. *WIREs Computattional Statistics*.

Sfetsos, A., & Siriopoulos, C. (2004). Time series forecasting with a hybrid clustering scheme and pattern recognition. *IEEE Transactions on Systems, Man, and Cybernetics, 34*(3), 399-405.

Shen, F., & Luo, N. (2016). *Investment Pattern Clustering Based on Online P2P Lending Platform.* Paper presented at the IEEE, Okayama, Japan.

Stetco, A., Zeng, X.-j., & Kaene, J. (2013). *Fuzzy cluster analysis of financial time series and their volatility assessment.* Paper presented at the IEEE International Conference on Systems, Man, and Cybernetics.

tslearn (Producer). (2017). tslearn's documentation. *tslearn's documentation*. Retrieved from https://tslearn.readthedocs.io/en/stable/

Ye, H., & Bellotti, A. (2019). Modelling Recovery Rates for Non-Performing Loans. *MDPI, Risks*.

BIOGRAPHY

Name-Surname	VAHREY SITSUKSAI
Academic Background	Bachelor's Degree in Business Administration with a Major in Finance from Mahidol University International College, Bangkok, Thailand in 2016 and Bachelor's Degree in Accountancy from University of the Thai Chamber of Commerce, Bangkok, Thailand in 2018
Experience	Student from the School of Statistics (Business Analytics and Data Science) from National Institute of Development Administration