ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

Analyzing Influential Factors on the Recovery Time of Non-Performing Loans: A Time Series and Machine Learning Approach

Vahrey Sitsuksai[1] and Ekarat Rattagan[2]

ABSTRACT

Non-performing Loans (NPLs) are critical factors that impede economic growth. Efficient management systems are required to expedite the resolution process for borrowers. However, managing NPLs remains challenging 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 balances, collateral appraisal values, and legal balance-to-appraisal values generally exhibit shorter recovery times. Additionally, the collateral subtype and the underwriting appraisal value of the collateral assets emerge as the most representative features of the clusters.

Article information:

Keywords: Recovery time, Non-performing Loans, Machine Learning, Time-series, K-means

Article history:
Received: July 22, 2023
Revised: September 14, 2023
Accepted: November 9, 2023
Published: December 2, 2023 (Online)

DOI: 10.37936/ecti-cit.2023174.253572

1. INTRODUCTION

During the Asian Financial Crisis (AFC) in 1997-87, Thailand witnessed a significant increase in the non-performing loans (NPLs) ratio, peaking at around 48% in 1999. This surge in NPLs jeopardized the banking system and impeded overall economic development. Consequently, the government implemented various policy measures to stabilize the financial system. These include temporary closure and subsequent consolidation of the financial institutions, along with establishing 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.

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, striving to sell their NPLs at the highest possible price. This contrasts with AMC's objective of purchasing at the lowest possible cost to maximize profitability. This dynamic often results in a mismatch in valuation pricing [7][8][16] and complicates the negotiation and transaction process between the two parties.

With these complexity, understanding the critical determinants of NPL valuation becomes crucial including the recovery rate [2][23] and the recovery time [7]. 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 Outstanding Principal Balance ("OPB") upon the loan acquisition. The recovery time is defined as the weighted average number of periods it takes the AMC to resolve the loan fully. The recovery time ranges between $[0, +\infty)$, and the infinity output is for

[1,2]The authors are with Graduate School of Applied Statistics, National Institute of Development Administration, Bangkok, Thailand, E-mail: vahrey.sit@stu.nida.ac.th and ekarat@as.nida.ac.th
[2]The corresponding author: ekarat@as.nida.ac.th

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 [2][6][23], develop models on repayment behavior [17], and examine factors that affect loan repayment capability [3]. Nevertheless, there is a notable gap in academic research, specifically focusing on the recovery time aspect of the loans, which could attribute 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.

In this research, we will analyze historical collection patterns, whi ch will later be used to compute recovery time by aggregating and normalizing the cash collections that the AMC received from each borrower in each period. It is important to note that the data used in this study only includes borrowers who have fully resolved. 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 called 'cash flow'. This term will be used consistently throughout the remainder of the paper to represent the combined cash collection values adjusted for normalization.

In this study, we aim to fill the gaps mentioned above, and the investigation seeks to provide the AMC with a deeper understanding of the factors influencing the recovery time and recovery pattern of the NPL through cash flow analysis. This could achieve by investigating the historical collection patterns and underlying factors affecting the cash flow by employing a combination of clustering and classification techniques. The cash flow data will serve as the primary input for conducting time-series clustering, with the Dynamic Time Warping (DTW) technique utilized as a method for calculating similarity distances between various cash flows. Upon obtaining the cluster of cash flow, these results will be incorporated as input data into a classification model to further explore the features contributing to the observed clustering patterns.

This insight from this study will enable the development of more effective loan recovery strategies and enhance the ability to identify borrower behavior patterns and trends associated with varying recovery times. Furthermore, this research has been structured as an initiative to apply machine learning methods to understand NPL recovery patterns. Utilizing this approach, along with a unique proprietary dataset of Thai AMC exposures, allows us to make valuable contributions to this field of research.

The remaining part of this paper is structured as follows. Section 2 provide the layout the fundamental knowledge requires to understand this paper including the algorithm and model applied in this paper and the basic knowledge about asset management business in Thailand. Section 3 describe the input data including both cash flow and demographic data for cash flow's analysis. Section 4 illustrates the methodology and the step-by-step procedure to conduct this research. Section 5 discusses the results of the study. Section 6 offers a conclusion.

2. BACKGROUND

2.1 Non-Performing Loans Recovery Process in Thailand

The AMC engages in the settlement process after acquiring an NPL portfolio from the selling bank. Settlement can be achieved through cash payment, entering into a troubled debt restructuring (TDR) scheme where the borrowers making payment in installment, or the transfer of collateral in the form of assets or equity, which are then restructured and sold in the market. In cases where the borrower refuses to settle, litigation is pursued, and the collateral is sold through auction to the third party buyer or bought by the AMC using the outstanding loan balance, resulting in its classification as NPA. Once the collateralized asset becomes an NPA, the AMC will sell the assets to potential buyers and eventually turn the assets into cash. The flowchart in Fig. 1 provides a visual representation of the AMC's business process and work flow after acquiring portfolio from the bank.

2.2 Cluster Analysis

Clustering is a powerful data mining technique 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 challenging to analyze using traditional methods. Clustering, as an unsupervised learning technique, helps to identify patterns [12] and structure in the data, providing a more profound insight 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

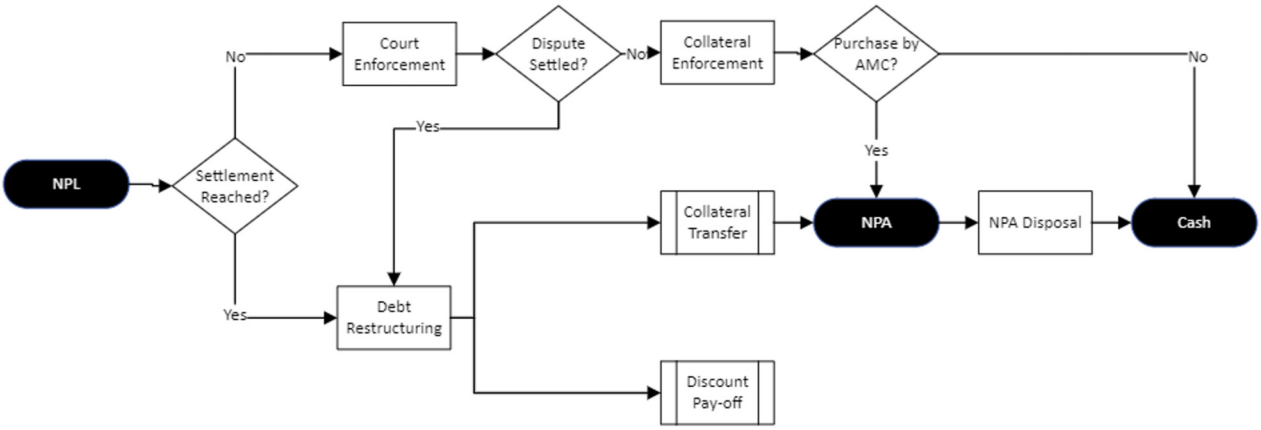


Fig.1: *NPL Management Process in Thailand.*

within clusters and maximize the distance between clusters [15].

2.2.1 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. It 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 .01, with -1 indicating that the point should belong to a different cluster and 1 indicating that it is correctly assigned to its cluster.

After selecting 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 [10].

2.2.2 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 and discovering an unusual and unexpected pattern from the time series data [13]. Pattern discovery is another study frequently observed using the time-series clustering method [9][20]. A hybrid technique could also lead to prediction and recommendation solutions [19].

Time-series is a sequence of continuous nominal values, and it comprises of 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 [11], and clustering on the entire object is a common method employed by several studies [1].

There are four components to construct a proper time-series clustering [1]. The first component is the time series representation, which transforms time series into the reduced dimensionality vector to a manageable size while preserving an essential character of the original data [18]. 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 [1]. If there are two-time series s and u consisting of T samples, each $s = (s_1, s_2, s_3, \ldots, s_t) \in S^T$ and $u = (u_1, u_2, u_3, \ldots, u_t) \in U^T$, 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 [5]. Below is the formula to calculate the distance under the DTW approach:

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

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

- it is a list of index pairs $\pi_k = [i_k, j_k]$ with $0 \leq i_k < n$ and $0 \leq j_k < m$
- π_0=(0,0) and π_K=$[n-1, m-1]$
- For all $k > 0, \pi_k = (i_k, j_k)$ is related to $\pi_(k-1) = (i_{k-1}, j_{k-1})$ as follows:
  - $i_{k-1} \leq i_k \leq i_{k-1} + 1$
  - $j_{k-1} \leq j_k \leq j_{k-1} + 1$
- n and m is the length of the first and second-time series, respectively

Fig. 2 presents the visualization of the distance calculation between two-time series for each approach.

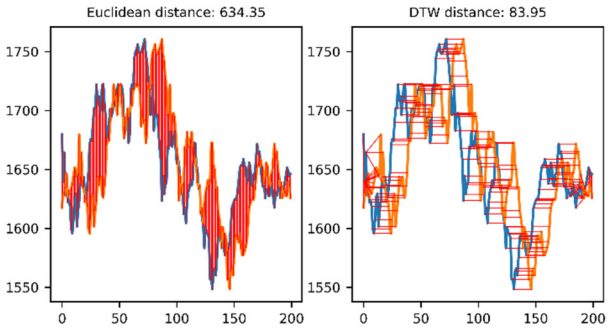


Fig.2: *Illustration of distance calculation applying different approaches.*

The third component is the cluster prototypes, 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.

2.3 Random Forest Classification Model

Employing a classification model to ascertain the key features that contribute to explaining the clustering results [4] 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 robust and widely adopted machine learning algorithm, the random forest model has demonstrated effectiveness in determining the most significant features influencing the result. Random forest operates by constructing many 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 [14]. The application of the random forest in this research expect to yield valuable insights into the factors that influence the recovery pattern of the NPL.

3. DATASET

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 2,839 cash flow profiles serving as input for the time-series clustering. Additionally, there are 50 unprocessed features that will be analyzed to determine the characteristics of the borrower. The information used in this research is divided into two sections, namely, data for time-series clustering and data for cluster analysis.

3.1 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 an 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 example of this is depicted in row 1, which showcases borrower 90's transaction. On 31 July 2015, this borrower made a significant payment amounting to 2.5 million baht. It is important to note 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 bor-

rower 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 Raw Data for Cluster Analysis

This set of information will comprise of the socio-demographic and loan file and collateral information of the borrower.

Table 2: *Sample of Socio-Demographic.*

Category	Features	Example of Data
Borrower Information	Borrower Type	• Corporate • Individual
	Industry	• Real Estate • Construction • Finance • Manufacturing
	Guaranteed	• Yes • No
	Legal Status	• Bankruptcy
	Residing Location	• Street • Province • Region • Zip code

The main socio-demographic features included in this database were the borrower type (individual vs. corporate), and operating industry (construction, finance, services, trading, etc).

Table 3: *Loan File and Collateral Data.*

Category	Features	Example of Data
Collateral Information	Collateral Type	• Commercial • Industry • Land • Residential • Others •
	Collateral Sub Type	• Agricultural Farming • Condominium • Single House • Detached House
	Type of Title Deed	• NSK3Kor • Chanod
	Land Area	XX. Squared Wah
	Building Area	XX. Squared Wah
	Street	Street name
	Province	Bangkok
	Region	• Central • North • South • West
Loan File	Legal Balance	8,000,000
	OPB	7,000,000
	Appraisal Value	8,500,000
	Loan-to-Value	121.4%

Loan file information refers to the outstanding loan value at the default event and its collateralized assets such as outstanding principal balance (OPB), legal balance (LB), and legal status of each borrower (defaulted, not under legal proceeding, filing, foreclosure, seized, liquidation, bankruptcy and etc). The collateral information includes number of collateralize assets, 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.

4. IMPLEMENTATION

0 delineates the implementation methodology adopted in this study, which comprises four integral processes. These processes encompass (1) data processing, (2) time-series clustering, (3) data classification for feature extraction, and finally, (4) data analysis. These collectively contribute to the robust and comprehensive exploration of the research subject.

4.1 Data Preprocessing

Data preprocessing was 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.

4.1.1 Data for Time Series Analysis

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 challenging to discover [22]. It is important to note that each borrower may have a different payment pattern, some might make a payment in lumpsum single amount, or making payment in installment. Hence, by doing the aggregation, the information for each 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 significant 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.

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

Where by : 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

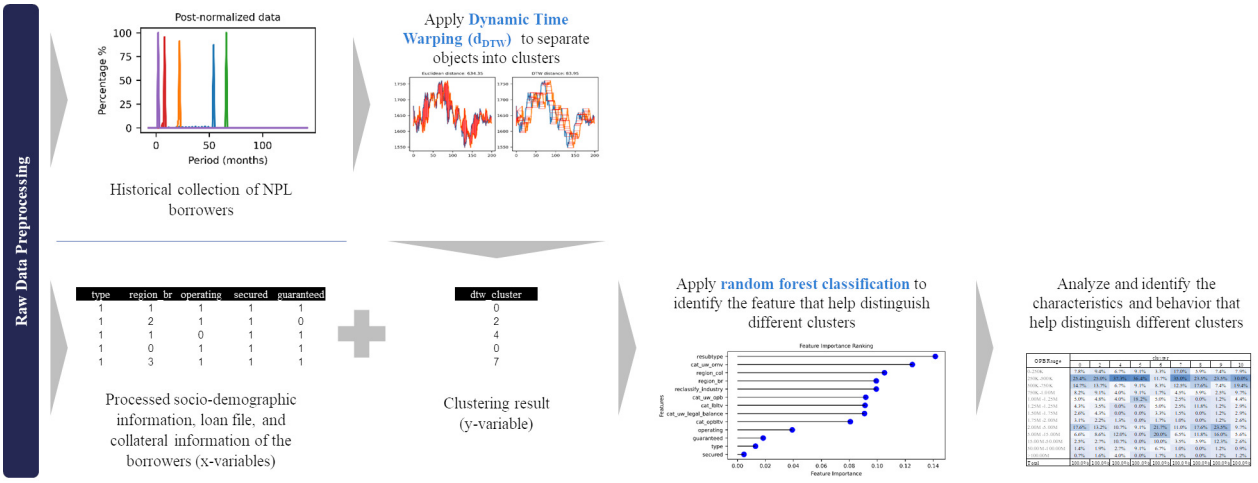


Fig.3: *Implementation Process.*

analysis. This led to a transformation of the data from Table 1 into an illustrative format that could be more easily analyzed in Fig. 4 below.

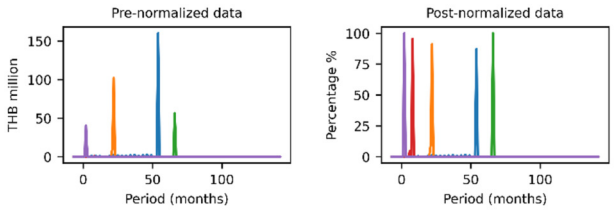


Fig.4: *Pre and Post-Normalized Time-Series Data.*

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 describing the collection timing. The chart 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.

4.1.2 Data for Cluster Analysis

The input data for the classification model application contains both continuous and categorical values. Continuous data primarily pertains to 'loan file' information, and due to the broad range of values in this dataset, the data has been grouped into categories to minimize distortion caused by outliers. Following this, a label encoder has been applied to the categorical data to ensure its suitability for the classification model, allowing it to perform optimally. Fig. 5 below illustrate the example of data that has been convert from continuous into categorical data.

The data prior to categorization was heavily skewed towards lower values. However, after being separated into categories, the data became more evenly distributed within each category.

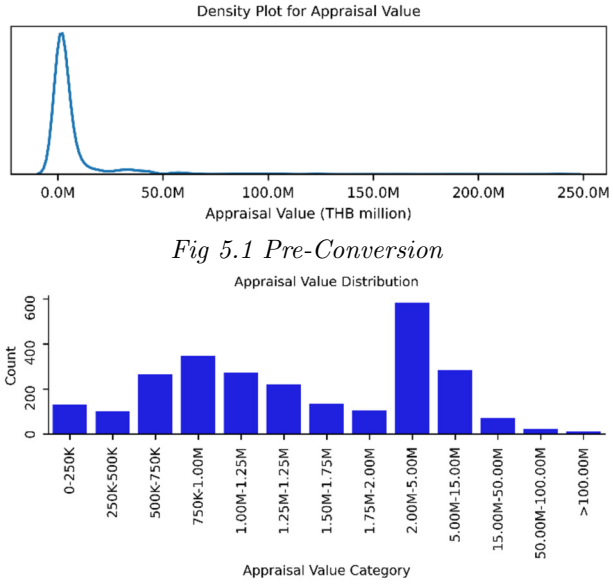


Fig.5: *Pre and Post-Appraisal Value Conversion.*

4.2 Applied Time-Series Clustering

As we sought to segment the time series data, the historical cash collection, into distinct clusters, the k-means clustering method was chosen as it is proficient at addressing clustering-type challenges. 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 our initial attempts, we applied the Euclidean distance for distance calculation; however, the results were not up to the mark, registering a negative silhouette score that ranged between -0.290 and -0.390. The result indicate 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. As the k-means approach requires a pre-determined number of clusters, we used the silhouette coefficient to determine the optimal number of clusters for further analysis.

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 calculation is visualized in the figure below.

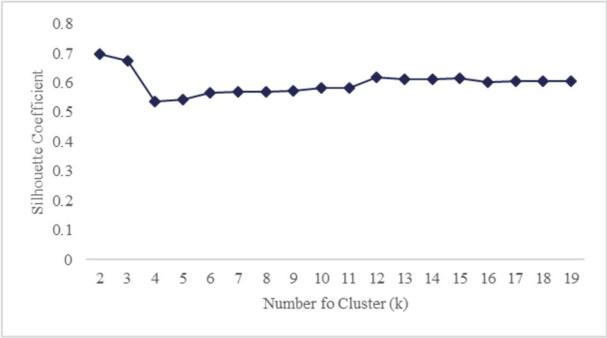


Fig.6: *Silhouette score from time-series clustering using DTW.*

Based on the Silhouette score, we considered the number of clusters to be 2, 3, 12, and 15. However, after further analysis, we decided to drop the analysis for k equal to 2 and 3 despite being selected based on the Silhouette score. This was due to the weighted average timing of the center clusters of these groups ranging between 4.755 months to the maximum of 23.494 months. This figure did not reflect reality as it is not possible for AMCs to have all the loans collected within the first two years after the acquisition. Hence, we only proceeded with k equal to 12 and 15 for the final analysis.

Fig. 7 above illustrates that slightly more than 50% of the members are grouped into cluster 0 based on their recovery patterns, with a weighted average recovery time of 23.88 months, or almost two years. As input data for the study only includes fully recovered loans, the summation of the cluster center should be more or less equal to 100 post-normalization, the summation that is less than 100 indicating that the total collections are lower than the expected result. With the significant deviation it could led to a distort of recovery time. For instance, with the same collection pattern and period but the summation of the cluster center is 50 instead of 100, it would result in a 50% faster recovery time because the weighted value is lower.

For k=12 clustering, the summation of the cluster center group 1, 3 and 11 is 64.88, 5.36 and 39.70, which significantly deviate from the expected value of 100. Therefore, these three groups have been exclude

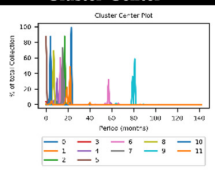
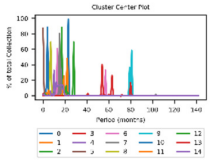
k	Cluster Center	Recovery Months			
12		group	member	month	center summation
		0	1428	23.889	99.618
		1	108	13.753	64.882
		2	373	17.070	98.587
		3	67	0.086	5.364
		4	75	13.015	88.140
		5	12	0.913	89.169
		6	60	30.650	102.082
		7	200	13.080	91.373
		8	47	7.598	99.056
		9	81	79.313	98.666
		10	343	4.777	97.986
		11	45	7.520	39.701
15		group	member	month	center summation
		0	1438	23.881	99.583
		1	103	13.167	62.414
		2	344	17.155	99.028
		3	63	0.048	2.454
		4	57	12.114	88.137
		5	12	0.913	89.169
		6	52	30.045	101.092
		7	178	15.304	92.092
		8	46	7.614	99.196
		9	78	79.390	98.729
		10	266	4.859	99.928
		11	37	7.815	48.069
		12	125	25.163	93.917
		13	24	58.401	93.872
		14	16	17.414	111.854

Fig.7: *Result of the time-series clustering using DTW.*

from the analysis.

Similarly, for k=15, we will exclude group 1, 3, and 11 from the analysis as the sum of normalized collection is 62.41, 2.45, and 48.07, respectively. For the other clusters, the summation of the cluster center ranges between 88.13 and 111.85, which is in acceptable range for further analysis.

As shown in Fig. 7, most line chart show a significant spike of 80% to 100% of the collection in a particular period. This is a result of recovery process of the AMC that encourage the borrower to either settle a large amount of debt in single or a few payments, or the collateral has been seized and sold entirely. It has been observed that the earlier spikes are usually a result of the borrower's willingness to settle, while the later spikes tend to indicate that the borrower has gone through the litigation process and their assets have been seized.

In some cases, there are patterns that show double or multiple spikes, which could imply that the borrower has more than one collateralized asset that was disposed of at different times. Alternatively, the borrower may have initially intended to settle but could not pay a sizable amount of cash at one time. It is worth noting that borrowers may enter thr TDR scheme, in which they make installment payments over a longer period of time, but this is relatively rare based on available information.

In conclusion, the clustering analysis of the cash flow has provided valuable insights into the repayment behavior of the borrowers. By examining the cluster centers and excluding those with significantly deviated sums of normalized collections, the study ensures a more accurate representation of the actual recovery patterns. This clustering analysis sets the foundation for further exploration of the factors influencing the recovery patterns, ultimately contributing

to a deeper understanding of the NPL recovery process and enabling the development of more effective loan recovery strategies.

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.

Nevertheless, before finalizing the application of the random forest model, several other traditional classification models, including logistic regression, gradient boosting classifier, k-nearest neighbor classifier, decision tree, and neural network classifier, were computed.

4.3 Feature Important Identification

Since the result of cluster 12 and 15 is quite similar, from this section onward, we will focus on analyzing the data from k=12 given a higher silhouette coefficient score.

Feature importance ranking offers insights into which factors most significantly influence the clustering results. This would further enable targeted analysis of select features to understand the distinct characteristics of each group within the cluster.

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 current sample size is tripled 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 performance evaluation results for these models were inferior to those of the random forest classifier. Consequently, the focus remained on the Random Forest model. Table 4. and Fig. 8 demonstrate the performance of the evaluation result.

Table 4: *Log-Loss Score of Classification Model.*

Classification Model	Log-loss score
Random Forest	0.310
Decision Tree	0.513
K-Nearest Neighbors	1.258
Gradient Boosting	1.231
Neural Network	1.553
Logistic Regression	1.607

An analysis of the feature importance results obtained from the model has been presented in Fig. 9, the higher score indicate that the features has higher influence to the classification model. From the result, 'collateral subtype' holds the highest importance (0.142), closely followed by 'appraisal value' (0.125). This indicates that the collateral subtype and the category of appraisal value significantly impact the recovery pattern of NPLs. The third most important feature is the region of the collateral (0.105), suggesting that the feature also plays a crucial role in determining the clustering results. Other features, such as 'OPB', 'percentage legal balance to value', 'operating industry' and 'residing location' of the borrowers, exhibit a moderate level of importance, ranging from approximately 0.10 to 0.11. The remaining features, including 'percentage of OPB to appraisal value', whether the borrowers are 'guaranteed' by others, and the 'type' of borrower, demonstrate a comparatively lower level of importance in the classification model. These results offer valuable insights into the key variables driving the recovery patterns, enabling the development of more effective loan recovery strategies and enhancing the ability to identify borrower behavior patterns and trends associated with different recovery times.

5. RESULTS AND DISCUSSION

Based on the cluster analysis, it appears that clusters 5, 8, and 10 have a collection span of less than one year, which could potentially align with the summation of the cluster center, which is lower than that of other clusters, ranging between approximately 89.169 and 97.986. A common theme among these three clusters is that the majority of the samples have a low legal balance-to-appraisal value ("cat_lbltv") with 27.3%, 31.3%, and 26.7% of the samples in the 25%-50% LTV range, respectively, while other clusters represent a more scattered LTV range. A distinguishing factor among these three clusters is the underwriting appraisal value ("cat_uw_omv"), with most of the samples in cluster 5 having a lower collateral value of around THB 0.75-1.0 million, while the majority of collateral assets in clusters 8 and 10 fall within the THB 1.0-1.25 million and THB 2.0-5.0 million buckets, respectively. This finding aligns with the business understanding that borrowers with lower LTVs tend to have faster collections, albeit with lower recovery rates. As the asset values are significantly higher than the loan amounts, borrowers may be more encouraged to pay down the loan rather than risk asset seizure by the AMC.

Clusters with longer recovery periods, such as clusters 6 and 9, have weighted average recovery times of 30.6 and 79.3 months, and summations of the cluster center ranging between 98.666 and 102.082. A common theme among these two clusters is that the majority of the samples fall within the OPB and appraisal value ranges of THB 2.0-5.0 million and THB 5.0-15.0 million buckets, with over a quarter of the le-

Fig.8: *Performance of the Classification Model.*

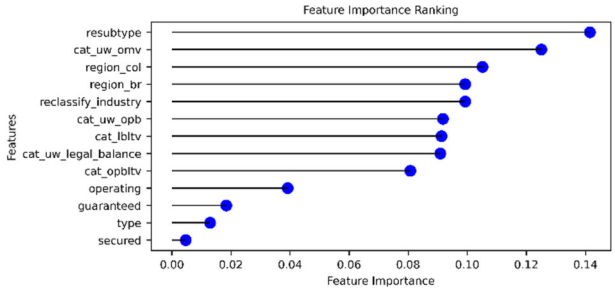


Fig.9: *Feature Importance Ranking based on Random Forest Model.*

gal balance-to-appraisal value range exceeding 200%. A distinguishing factor between these two clusters lies in the collateral subtype. Most of samples in cluster 6 are concentrated in the combination bucket, where borrowers have more than two pieces of collateral (23.3%), as well as land and building (18.3%). In contrast, cluster 9 does not exhibit a significant trend but is characterized by assets scattered across various collateral subtypes, unlike other clusters that show concentration in "single-house" properties.

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 quicker 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 cautiously approached 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.

Collateral Sub Type	cluster								
	0	2	4	5	6	7	8	9	10
single house	33.2%	40.6%	44.0%	36.4%	25.0%	44.0%	47.1%	22.2%	39.4%
townhouse	16.7%	10.2%	5.3%	9.1%	6.7%	12.0%	5.9%	9.9%	15.0%
vacant land	8.0%	8.9%	5.3%	0.0%	6.7%	9.0%	5.9%	3.7%	12.9%
shophouse	8.1%	7.8%	4.0%	45.5%	1.7%	6.5%	0.0%	11.1%	10.0%
land and building	9.6%	5.9%	8.0%	0.0%	18.3%	5.5%	5.9%	12.3%	2.4%
land	5.2%	5.6%	12.0%	0.0%	8.3%	6.0%	17.6%	8.6%	2.4%
detached house	5.0%	7.0%	2.7%	0.0%	1.7%	5.5%	0.0%	6.2%	3.5%
combination	2.3%	4.3%	8.0%	0.0%	23.3%	5.0%	11.8%	13.6%	3.2%
condominium	4.1%	3.0%	0.0%	0.0%	3.3%	3.5%	5.9%	3.7%	3.2%
others	2.1%	2.7%	1.3%	0.0%	1.7%	1.0%	0.0%	1.2%	1.5%
agricultural	1.6%	1.9%	5.3%	9.1%	3.3%	1.0%	0.0%	2.5%	2.1%
factory/warehouse	1.7%	1.3%	1.3%	0.0%	0.0%	0.5%	0.0%	1.2%	1.2%
land and single house	0.9%	0.5%	2.7%	0.0%	0.0%	0.5%	0.0%	1.2%	2.4%
semi-detached house	1.1%	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	1.2%	0.6%
commercial	0.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1.2%	0.3%
Total	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

Fig.10: *Collateral Sub Type by Cluster.*

Appraisal Value Range	cluster								
	0	2	4	5	6	7	8	9	10
0-250K	5.7%	3.5%	8.0%	9.1%	11.7%	5.5%	5.9%	6.2%	2.1%
250K-500K	3.4%	3.8%	4.0%	9.1%	0.0%	5.5%	5.9%	6.2%	5.3%
500K-750K	10.8%	9.1%	6.7%	9.1%	3.3%	17.0%	5.9%	7.4%	9.1%
750K-1.00M	13.9%	12.9%	16.0%	27.3%	5.0%	13.0%	11.8%	8.6%	15.0%
1.00M-1.25M	11.0%	9.9%	10.7%	9.1%	8.3%	13.0%	17.6%	3.7%	10.9%
1.25M-1.25M	8.8%	10.8%	8.0%	9.1%	3.3%	9.5%	0.0%	4.9%	7.6%
1.50M-1.75M	4.6%	5.1%	1.3%	0.0%	8.3%	2.0%	11.8%	4.9%	11.2%
1.75M-2.00M	4.3%	4.8%	4.0%	9.1%	1.7%	2.5%	5.9%	2.5%	4.7%
2.00M-5.00M	24.7%	22.3%	17.3%	9.1%	21.7%	17.0%	11.8%	23.5%	21.5%
5.00M-15.00M	9.9%	12.6%	12.0%	0.0%	30.0%	9.5%	17.6%	21.0%	9.7%
15.00M-50.00M	2.1%	2.7%	9.3%	0.0%	3.3%	4.0%	5.9%	8.6%	2.1%
50.00M-100.00M	0.6%	1.9%	2.7%	9.1%	1.7%	0.5%	0.0%	2.5%	0.6%
>100.00M	0.4%	0.5%	0.0%	0.0%	1.7%	1.0%	0.0%	0.0%	0.3%
Total	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

Fig.11: *Asset Appraisal Value by Cluster.*

OPB Range	cluster								
	0	2	4	5	6	7	8	9	10
0-250K	7.8%	9.4%	6.7%	9.1%	3.3%	17.0%	5.9%	7.4%	7.9%
250K-500K	25.4%	25.0%	37.3%	36.4%	11.7%	35.0%	23.5%	23.5%	30.0%
500K-750K	14.7%	13.7%	6.7%	9.1%	8.3%	12.5%	17.6%	7.4%	19.4%
750K-1.00M	8.2%	9.1%	4.0%	9.1%	1.7%	4.5%	5.9%	2.5%	9.7%
1.00M-1.25M	5.0%	4.8%	4.0%	18.2%	5.0%	2.5%	0.0%	1.2%	4.4%
1.25M-1.25M	4.3%	3.5%	0.0%	0.0%	5.0%	2.5%	11.8%	1.2%	2.9%
1.50M-1.75M	2.6%	4.3%	0.0%	0.0%	3.3%	1.5%	0.0%	1.2%	2.9%
1.75M-2.00M	3.1%	2.2%	1.3%	0.0%	1.7%	1.0%	0.0%	1.2%	2.6%
2.00M-5.00M	17.6%	13.2%	10.7%	9.1%	21.7%	11.0%	17.6%	23.5%	9.7%
5.00M-15.00M	6.6%	8.6%	12.0%	0.0%	20.0%	6.5%	11.8%	16.0%	5.6%
15.00M-50.00M	2.5%	2.7%	10.7%	0.0%	10.0%	3.5%	5.9%	12.3%	2.6%
50.00M-100.00M	1.4%	1.9%	2.7%	9.1%	6.7%	1.0%	0.0%	1.2%	0.9%
>100.00M	0.7%	1.6%	4.0%	0.0%	1.7%	1.5%	0.0%	1.2%	1.2%
Total	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

Fig.12: *Outstanding Principal Balance Range by Cluster.*

Legal Balance-to-Appraisal Value	cluster								
	0	2	4	5	6	7	8	9	10
0-25%	7.9%	12.3%	7.4%	0.0%	3.8%	12.7%	6.3%	10.5%	12.4%
25%-50%	20.2%	20.9%	33.8%	27.3%	22.6%	34.9%	31.3%	14.5%	26.7%
50%-75%	17.9%	20.1%	17.6%	9.1%	1.9%	23.3%	12.5%	15.8%	22.1%
75%-100%	19.2%	13.6%	14.7%	18.2%	22.6%	12.2%	6.3%	13.2%	16.4%
100%-125%	11.5%	10.6%	2.9%	9.1%	11.3%	3.7%	12.5%	6.6%	10.0%
125%-150%	6.6%	5.3%	1.5%	27.3%	3.8%	1.1%	12.5%	9.2%	3.6%
150%-175%	3.6%	3.1%	1.5%	9.1%	5.7%	3.2%	6.3%	2.6%	2.7%
175%-200%	1.7%	1.7%	2.9%	0.0%	1.9%	1.1%	6.3%	2.6%	0.9%
>200%	11.4%	12.5%	17.6%	0.0%	26.4%	7.9%	6.3%	25.0%	5.2%
Total	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

Fig.13: *Legal Balance-to-Appraisal Value Range by Cluster.*

6. CONCLUSIONS

This study aims to develop an understanding of forecasting recovery time 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 impacting the clustering results. However, some clusters were eliminated from the analysis due to unrealistic values.

Based on the feature importance ranking, collateral subtype and underwriting appraisal value of the collateral assets are the features that most represent the clusters. However, given the portfolio's high concentration in "single-house" properties, it is challenging to analyze based on that aspect alone. Therefore, a combination of four features, including collateral subtype, appraisal value, OPB, and legal balance-to-appraisal value, was used to extract meaningful insights.

The study demonstrates that borrowers with lower OPB, appraisal value, and legal balance-to-appraisal value generally exhibit shorter recovery times. This knowledge can be leveraged to refine underwriting model, develop more effective loan recovery strategies and enhance the ability to identify borrower behavior patterns and trends associated with varying recovery times.

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.

References

[1] S. Aghabozorgi, A. S. Shirkhorshidi and T. Y. Wah, "Time-series clustering – A decade review," *Information Systems*, vol. 53, pp. 16-38, 2015.

[2] A. Bellotti, D. Brigo, P. Gambetti and F. Vrins, "Forecasting Recovery Rates on Non-Performing Loans with Machine Learning," *International Journal of Forecasting*, vol. 37, no. 1, pp.428-444, 2021.

[3] N. Bhatt and S.-Y. Tang, "Determinants of Repayment in Microcredit: Evidence from Pro-

grams in the United States," *International Journal of Urban and Regional Research*, vol. 26, no. 2, pp. 360-376, 2002.
[4] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5-32, 2001.
[5] C. Cassisi, P. Montalto, M. Aliotta, A. Cannata and A. Pulvirenti, "Similarity measures and dimensionality reduction technique for time series data mining," *Advances in Data Mining Knowledge Discovery and Applications.*, InTech, pp. 70-96, Sep. 2012.
[6] D. Cheng and P. Cirillo, "A Reinforced Urn Process Modeling of Recovery Rates and Recovery Times," *Journal of Banking & Finance*, vol. 96, pp. 1-17, Nov. 2018.
[7] Ciavoliello, L. G., et al. (2016). What's the value of NPLs., Banca Di Italia.
[8] J. Fell, M. Grodzicki, R. Martin and E. O'Brien, "A Role for Systemic Asset Management Companies in Solving Europe's Non-Performing Loan Problems," *European Economy Banks*, pp. 71-85, 2017.
[9] F. Iglesias and W. Kastner, "Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns," *Energies*, vol. 6, no. 2, pp. 579–597, Jan. 2013.
[10] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, New Jersey, Prentice-Hall, 1948, pp. 96-97.
[11] R. P. Kumar and P. Nagabhushan, "Time Series as a Point – A Novel Approach for Time Series Cluster Visualization," *International Conference on Data Mining*, 2006.
[12] Y. Lei *et al.*, "Ground truth bias in external cluster validity indices," *Pattern recognition*, vol. 65, pp. 58-70, 2017.
[13] M. Leng, X. Lai, G. Tan and X. Xu, "Time series representation for anomaly detection," *2009 2nd IEEE International Conference on Computer Science and Information Technology*, Beijing, pp. 628-632, 2009.
[14] A. Liaw and M. Wiener, "Classification and Regression by RandomForest," *R News*, vol. 2, no. 3, pp. 18-22, Dec. 2002.
[15] R. Ma and R. Angryk, "Distance and Density Clustering for Time Series Data," *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, New Orleans, LA, USA, pp. 25-32, 2017.
[16] F. Pauer and S. Pichler, "Sell or Hold? On the Value of Non-Performing Loans and Mandatory Write-Off Rules," *SSRN*, 2021.
[17] J. Paxton, D. Graham and C. Thraen, "Modeling Group Loan Repayment Behavior: New Insights from Burkina Faso," *Economic Development and Cultural Change*, vol. 48, no. 3, pp. 639-655, 2000.
[18] S. J. Wilson, "Data representation for time series data mining: time domain approaches," *WIREs Computational Statistics*, vol. 9: e1392, 2017.
[19] A. Sfetsos and C. Siriopoulos, "Time series forecasting with a hybrid clustering scheme and pattern recognition," in *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 34, no. 3, pp. 399-405, May 2004.
[20] F. Shen and N. Luo, "Investment pattern clustering based on online P2P lending platform," *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, Okayama, Japan, pp. 1-6, 2016.
[21] Shutaywi, M.; Kachouie, N.N. Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering. Entropy 2021, 23, 759. https://doi.org/10.3390/e2306075
[22] A. Stetco, X. -j. Zeng and J. Keane, "Fuzzy Cluster Analysis of Financial Time Series and Their Volatility Assessment," *2013 IEEE International Conference on Systems, Man, and Cybernetics*, Manchester, UK, pp. 91-96, 2013.
[23] H. Ye and A. Bellotti, "Modelling Recovery Rates for Non-Performing Loans," *Risks*, vol. 7, no. 1, p. 19, Feb. 2019.

Vahrey Sitsuksai received a Bachelor of Business Administration from Mahidol University International College (MUIC) in 2016, and a Bachelor of Accountancy from University of the Thai Chamber of Commerce (UTCC) in 2018, and currently pursuing a Master's Degree in Business Analytics and Data Science at the Graduate School of Applied Statistics, National Institute of Development Administration (NIDA) in Bangkok, Thailand.

Ekarat Rattagan received a B.Arch from Chulalongkorn University (CU) in 1999, an M.Sc. in Information Technology from King Mongkut's University of Technology Thonburi (KMUTT) in Bangkok, Thailand, in 2003, and a Ph.D. in Electrical Engineering and Computer Science from National Chiao Tung University (NCTU) in Taiwan in 2016. During his Ph.D., he was fortunate to be advised by Prof. Ying-Dar Lin. Currently, he is a lecturer and the head of the M.Sc. Program in Data Analytics and Data Science (DADS) at the Graduate School of Applied Statistics, National Institute of Development Administration (NIDA) in Bangkok, Thailand.