Application of Clustering in the Dimensionality Reduction Algorithms for Separation of Financial Status of Commercial Banks in Ukraine

Halyna Velykoivanenko*, Vladyslav Korchynskyi

Department of Mathematical Modeling and Statistics, Institute Information Technologies in Economics, Kyiv National Economic University named after Vadym Hetman, 03057, Kyiv, Ukraine

Received October 20, 2021; Revised November 30, 2021; Accepted December 22, 2021

Cite This Paper in the following Citation Styles

(a): [1] Halyna Velykoivanenko, Vladyslav Korchynskyi, "Application of Clustering in the Dimensionality Reduction Algorithms for Separation of Financial Status of Commercial Banks in Ukraine," *Universal Journal of Accounting and Finance*, Vol. 10, No. 1, pp. 148-160, 2022. DOI: 10.13189/ujaf.2022.100116.

(b): Halyna Velykoivanenko, Vladyslav Korchynskyi (2022). *Application of Clustering in the Dimensionality Reduction Algorithms for Separation of Financial Status of Commercial Banks in Ukraine*. *Universal Journal of Accounting and Finance*, 10(1), 148-160. DOI: 10.13189/ujaf.2022.100116.

Copyright©2022 by authors, all rights reserved. Authors agree that this article remains permanently open access under the terms of the Creative Commons Attribution License 4.0 International License

Abstract The issue of determining the financial condition of commercial banks and separating investment-attractive banks from problem banks on this ground is extremely important for developing countries. The aim of this study is to make sure on the example of Ukraine that commercial banks really form separate clusters, where more reliable, stable and efficient banks are well separable from less successful ones in this regard. The study used the t-SNE and UMAP dimensionality reduction algorithms, and the Ward's Agglomerative Hierarchical Clustering algorithm. The results of visual analysis of two-dimensional t-SNE projections show that banks of different degrees of risk are well separable and have their own specifics. Clustering in the UMAP algorithm allowed distinguishing clusters with banks of Class A, "mid-tier" and problematic banks by different parameters. The t-SNE and UMAP algorithms for solving the problem are compared. The results show that a purely visual analysis of the two-dimensional map for the banks over the last period is best made using the t-SNE algorithm. UMAP, on the other hand, is proved to be excellent when used in tandem with the clustering algorithm.

Keywords Banks, Financial Condition, Machine Learning, Clustering, Dimensionality Reduction, Ukraine

1. Introduction

The financial condition of commercial banks is one of the most important categories of the financial sector of any developed or developing country. The financial performance of banking institutions significantly affects the financial stability of the entire economy [1]. But an important and difficult issue is the formation of appropriate tools for assessing the financial condition of banks. Different techniques can give different results. The assessment of the reliability of the financial system in the country as a whole will depend on this. The reaction and behaviour of banking institutions during the 2007-2008 crises have been examined and analysed over the past decade by researchers around the world in order to assess the effect, the resulting damage and/or develop methods to mitigate future financial impacts [2-5]. This issue has become particularly important for developing countries. Such countries have weak financial systems and the reliability of the banking system is important. Confidence in the banking system provides confidence in the economy as a whole [5].

First of all, let's consider the concept of financial condition. We can determine the financial condition of a commercial bank as a comprehensive economic characteristic of its activities in the short, medium and long term, determined by the structure of the bank's

sources of funds and the effectiveness of their placement. The financial condition of a commercial bank determines a number of various external and internal factors, individual characteristics of the bank, which include the quality of assets, capital adequacy, loan to assets ratio, and the quality of management [6]. Sometimes you can come across another term, which is, however, very close by its meaning — financial performance. Studies of financial performance of banks also involve such characteristics of banking as capital adequacy, asset quality, quality of management, as well as profitability and bank liquidity [7].

We are going to consider the mentioned categories in the context of the Ukrainian financial market, and it is necessary to note some of its features. It is relatively young and is still at the stage of active development [8], it is floating to some extent, being in a state of uncertainty, in the process of transition from the post-Soviet model to European standards [9,10].

Despite the fact that this transition can be called relatively successful, there are still problematic points in the banking sector of Ukraine that do not allow asserting the unequivocal reliability and efficiency of commercial banks in the system: a high proportion of low-quality assets, a high level of credit risk in general, low efficiency of some systemically important banks, a decrease in the return on assets and capital and the instability of these indicators over time.

Based on the above, it can be unequivocally stated that the question of identifying the most effective, reliable, stable banks is urgent for individuals and legal entities, for whom it is important to determine the investment attractiveness of banks for the correct investment of their funds, which indicates the topicality of this research.

To conduct the study, we decided to move away from classical simulation approaches and turn to more innovative tools and methods of unsupervised machine learning for data dimensionality reduction algorithms, since we assume that they enclose unrevealed potential for solving problems in the domain under study.

There are many works in the field of data clustering in the scientific field, and a significant part of them deal with the development of the algorithms themselves, not to their application, for example, the work of researchers [11]. The studies in the banking field often focus on a particular side of the bank, the business line, for example author [12] examined the bank's credit risk. The subject of a comprehensive analysis of the banking system using methods and tools of unsupervised machine learning with data dimensionality reduction algorithms has not been widely covered in the scientific literature, while it is even more difficult to find and qualitative research in the context of Ukrainian realities. Therefore, we can argue that this study represents scientific novelty, and is aimed at filling the above research gap.

Two objectives were identified that are aimed at

achieving the goal of this study. The primary objective is to make sure that the separation of commercial banks (their clustering), based on the indicators of different aspects of their activities, is possible, and that successful investment-attractive banks are separable from problem banks, since banks with similar indicators in the multidimensional space are next to each other. Secondary objective is to verify whether t-SNE is a suitable dimensionality reduction algorithm in the context of working with data in a banking domain, or it is good for data visualization only and UMAP suits better.

The rest of the article is structured as follows. Literature Review — contains a review of the world scientific literature covering various aspects of this research, a list of advanced hypotheses. Material and Methods — contains a detailed description of the data used in this work and the approach to their pre-processing, the used dimensionality reduction algorithms and unsupervised machine learning. The Results and Discussion section provides the proposed approaches to diagnosing the financial condition of Ukrainian commercial banks through the t-SNE and UMAP dimensionality reduction algorithms, as well as the Ward's Agglomerative Hierarchical Clustering algorithm. Research hypotheses have been worked out. The last part is Conclusions, which additionally consider the limitations of this work and suggest directions for further research.

2. Literature Review

One of the key categories of this work is the financial condition of a commercial bank. Some works, for example, author [13], say that indicators of banks' financial performance include profitability and return on assets. But this approach is more focused on bank owners than on their customers. From the point of view of clients, more important parameters of the bank's financial stability are its liquidity and stability than profitability [13]. This category is also often referred to in scientific literature as financial performance, and has been considered by researchers in different contexts, for example, in the context of corporate social responsibility [14]. We consider the financial performance of the bank as a result of the effect of different indicators of its operation. A similar approach can be found in researchers [7], where the category of financial performance itself and the factors influencing it are considered.

If we consider the economic and mathematical models for diagnosing the financial condition of commercial banks, the frequent use of classical methods can be noted. Researcher [15] considers a model based on multiple discriminant analysis and compares it with logistic regression in the context of diagnosing Slovak bank bankruptcies. As a result of the analysis, the author comes to the conclusion that “MDA is not a recommended

method for bankruptcy prediction”, and “in terms of logit model estimation, we have to admit its uselessness” [15].

Machine learning methods can be distinguished among more modern ones. For example, some authors consider machine learning based on modified logistic regression, which has shown good results in solving the problem of credit scoring. Ensemble models show high efficiency in the tasks of analysing bank performance, for example, researchers [16] consider the bagging method built on three tree-based models, which is used to predict the key factors of profitability of Turkish banks (Development and Investment Banks).

One of the types of machine learning is unsupervised machine learning, models of this type solve clustering problems. Clustering has been used in the research of many scientists; it is advisable to familiarize with it as a separate type of algorithms on the basis of works dealing with the review and comparative analysis of such algorithms [17, 18]. Central banks around the world develop and implement their own systems and algorithms that best meet the characteristics of national financial systems. At the international level, such systems are still being agreed for comparability. Clustering is a popular method because it allows you to divide banks into different groups for which there are common features. This approach provides an opportunity to develop recommendations for each group of banks, taking into account the problems and risks that are specific to this group [17].

It is difficult to argue that the clustering of commercial banks has received much attention in the scientific literature. The article by researcher [19] considers the problem of identifying failing banks using neural networks such as self-organizing maps, concluding that “the results show that self-organizing maps can successfully carry out bank clustering tasks and identify banks that require immediate attention from the regulatory bodies” [19]. The work of researcher [20] deals with the clustering of Ukrainian banks using the Ward agglomerative hierarchical clustering, with the aim of dividing them into three groups (“Growth”, “Stabilization”, “Decline”). Researchers [21] also applied hierarchical cluster analysis to identify and separate problem banks from the rest based on 12 performance indicators. Considering that there is a certain research gap in studies of the Ukrainian financial market using machine learning and clustering methods, and that many existing works may have lost their relevance due to the high dynamism of the financial sector, we would like to fill this gap in this work and investigate the following research questions:

1. Separation of Ukrainian commercial banks (their clustering) based on a wide range of indicators of their activities is possible, and effective investment-attractive banks are separable from problem banks.

2. Problem banks tend to fall into areas where closed banks were in the last period before they were liquidated, and this can be used to identify them even before clustering is applied.

An important stage in our research is the analysis of multidimensional data, which is a very difficult, sometimes impossible task due to such a phenomenon as the “curse of dimensionality” [22]. Moreover, it is necessary for the visual analysis to display a multivariate array of model features in two- or three-dimensional dimensionalities. Dimensionality reduction algorithms are used to solve this problem. The t-Stochastic Neighbour Embedding (t-SNE) algorithm is one of the most popular among them, it was proposed researchers [23] for multivariate data visualization. However, some researchers [24] see this algorithm as a pre-processor for further clustering, when clustering is performed on a two-dimensional map obtained after dimensionality reduction. Other researchers [25, 26], on the contrary, point to the limitations of the algorithm, which include a strong dependence on the local data structure and the inability to preserve the global structure, the inability to allocate outliers, and high requirements for computing power. The Uniform Manifold Approximation and Projection (UMAP) algorithm proposed by can be used as an alternative. It is relatively young, but has already gained popularity in scientific research in the fields of bioinformatics, materials science, natural sciences, machine learning. A comprehensive literature review on this topic is provided by researchers [27, 28] noted the computational efficiency of the algorithm (speed of operation), the ability to form meaningful clusters, and high reproducibility. Researchers [29] test four dimensionality reduction algorithms on 71 genetic datasets, stating that UMAP is superior to PCA and Multidimensional Scaling (MDS) algorithms, and shows advantages over t-SNE.

However, the work of Kobak & Berens (2019), dealing with a detailed analysis of the t-SNE algorithm and how to optimize it, also compared t-SNE and UMAP and concluded that, with the correct settings of the t-SNE parameters, it allows maintaining a global structure no worse than UMAP, you can achieve reproducible results when using initialization with the help of PCA, and the Fit-SNE modification can significantly increase the speed of the algorithm.

So, it is difficult to speak for sure about the superiority of any of the algorithms, and given that there is a certain polarity of opinions in the research field, we would like to put forward and test the following hypothesis:

H3: The UMAP algorithm is the most suitable for data dimensionality reduction in the study of Ukrainian commercial banks from the point of economic interpretation.

3. Research Methodology

The research design of this study includes all the stages necessary to accomplish its key objectives: Data collection, Feature preparation & selection, data analysis & pre-processing, Dimensionality reduction, Clustering.

3.1. Data Collection

The object of our research is Ukrainian commercial banks, and we collected the relevant data of the NBU supervisory statistics for its conduct, namely, quarterly data of the grouped balance sheet balances of banks for the period QI 2015 to QIV 2020, broken down into four groups: Assets, Liabilities, Equity, Financial results. (NBU data, Aggregated outstanding amounts on balance sheet accounts of the Ukrainian banks, 'Aggregation_eng.zip').

3.2. Feature Preparation & Selection, Data Analysis & Pre-processing

Taking into account the specifics of our research, we relied on a number of Ukrainian-language sources which provide an overview of diverse banking indicators. As a result, we identified 143 coefficients, including 63 ones that can be calculated based on data from open statistics of the NBU.

We have taken a number of measures to eliminate distortions in the data. First, the N/A data (110 observations out of 2243, i.e. 5% of the dataset) were replaced with 0 with the intention of keeping some banks in the sample, but giving all such observations a similarity characteristic. We assume that zero values of the ratios more often indicate the problem of banks.

Second, we applied a data winsorization approach to limit outliers [30]. We additionally used the classic standardization to smooth the distribution, where X – raw score; μ – population mean; σ – population standard deviation.

We also used robust scaler paired with winsorization. Robust scaler scales features using statistics that are robust to outliers. This method removes the median and scales the data in the range between 1st quartile and 3rd quartile. i.e., in between 25th quantile and 75th quantile range (interquartile range). Therefore, its formula is as follows:

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

We obtained an improvement in the quality of the dataset as a result of that data processing. The total mathematical expectation across the entire dataset decreased from 475.53 to 0.00 at the standard scaler and to 0.30 at the robust scaler; the overall standard deviation decreased from 10394.29 to 1.00 and 2.08, respectively.

$$\frac{x_i - Q_1(x)}{Q_3(x) - Q_1(x)} \quad (2)$$

Third, we performed multicollinearity analysis using Pearson's pairwise correlation and Variance Inflation Factor (VIF) metrics, relying on the work of researchers [31], in which these methods, among others, are considered as tools for determining the correlation between factors, and their thresholds are indicated at 0.7 and higher for the correlation and 10 or more for the VIF indicator. It is known that multicollinearity is extremely difficult and often impossible to correct by changing the data itself. But the solution to this problem is the feature selection process. We gradually added features to the model following a bottom-up approach based on the pairwise correlation metrics, VIF and two complementary methods, Mean Absolute Difference [32] and Spectral Feature Selection (SPEC) [33]. We also relied on our expertise in the banking area, primarily testing the indicators that, in our opinion, were the most important for inclusion in the model for assessing banking performance. A total of 17 features are included in the model. A complete list is provided in the Appendix. It was decided to leave one pair of coefficients (K12 and K113), in which the pair correlation is higher than 0.7 (0.98) and the VIF is greater than 10 (397.61 and 438.49, respectively), since their inclusion in the model has a positive effect on the interpretability and quality of data separation in two-dimensional space. In this case, this pair of features can be considered as one feature with increased weight in the model. As a result of the selection of indicators, the average VIF value for all data decreased from 1010.140 to 56.096. A further change in the number of features and/or their replacement did not affect the quality of clustering and the interpretability of the results at best; it had a negative effect at worst.

3.3. Dimensionality Reduction

We opted for two algorithms to reduce the dimensionality of multidimensional data: t-SNE and UMAP. We also considered the Principal Component Analysis (PCA) algorithm, but it showed significantly worse results, and the quality of clustering and data interpretability in the two-dimensional plane turned out to be unacceptably low.

When working with algorithms and setting their hyper-parameters, we relied on the original work of the authors of UMAP [34] and t-SNE [23], as well as on the work of researchers [29, 35].

3.4. Clustering

We have chosen the Ward Agglomerative Hierarchical Clustering algorithm to solve the problem of data clustering. Hierarchical clustering is a cluster analysis technique that aims to build a hierarchy of objects. Its overview can be found in researchers [36, 37]. The latter also provides an implementation of the algorithm in

To assess the quality of clustering, 3 metrics were used that are suitable for assessing when the base true marks of the clusters are unknown: the Silhouette coefficient [39], Calinski-Harabasz score [40] and Davies-Bouldin score [41].

All other stages of the research were carried out using the Python programming language in the Google Collab environment, with the support of the freely available libraries `numpy`, `scipy`, `pandas`, `matplotlib`, `seaborn`, `sklearn`, `statsmodel`, `umap`, `pyclusterend`.

In our research, we came to several approaches to diagnosing the financial condition of Ukrainian commercial banks.

In the first approach, we conduct a purely visual analysis of the distribution of points, using only dimensionality reduction algorithms, mapping the multidimensional space (these are 17 dimensions in our case) in a two-dimensional projection. To do this, we use the t-SNE algorithm, which has shown a fairly good result in preserving the local and global data structure and excellent results for subsequent visual interpretation. Our main goal is to find out if there are any regularities and economic sense in the distribution of banks on a two-dimensional projection in the context of one period (the last one). Let us refer to Figure 1 for this purpose.

Figure 1. Analysis of the distribution of bank points on a two-dimensional projection map for QIV 2020 (or for the last available period for inactive banks)

On the map in Figure 1, yellow labels indicate active banks, and red ones — those that have left the market (data for the last available period). If we consider only active banks (that is, yellow labels on the projection map), we can clearly see that all banks are divided into two sectors: the banks in the lower part are clearly separated by a space from the banks in the upper part of the map. We will draw R1 through this gap to better visually indicate the separation. We see it necessary to emphasize, that this line is drawn not based on any specific calculations but on a visual analysis only and it has a single purpose to highlight this gap between active banks (yellow labels) that could be noticed in the middle of the projection map. We can consider this R1 line simply as a visual separator that divides the whole map into two major groups: above and below the line. We can also notice that the bank points form groups among themselves. Let's analyse the separated areas and banks therein, relying on two approaches: considering the rankings of banks (credit ratings of national ranking agencies, rankings of the reputable Ukrainian financial Internet publication Minfin and the official ranking of the NBU), and analysing a number of key ratios (K13. Credit losses provision ratio; K113. Credit-investment portfolio share; K51. Highly liquid assets share; K31. Autonomy ratio (capital adequacy); K68. Return on costs).

Having considered all the banks, we can draw the following conclusions. Based on our visual analysis of the above projection map we can distinguish five major groups of banks. The area A+ (the lower sector, the bank points converge to the lower left corner) grouped first-class banks, having the highest rankings of other experts and showing good results in terms of their indicators. There are the TOP-10 banks according to the Minfin ranking, 7 out of 10 TOP-10 banks according to the NBU ranking, 11 out of 12 banks have the highest ua AAA ranking. The areas B+ (the lower sector, centre) are mid-tier banks, mostly small businesses with above-average performance. The Corp area (lower sector, on the right) is predominantly corporate, international and/or investment banks. The upper part of the upper sector grouped state-owned banks, which are rather problematic in terms of their indicators; however, they cannot be unequivocally classified as such given the support of the state. The rest of the banks above R1 can be attributed to the Risk group: these are potentially problem banks, they form their own separate subgroups, each being characterized by its own set of indicators, according to which banks can be classified as problem banks. As Figure 1 also shows, most of the closed banks are located above the straight line, in the same place as problem banks (59 out of 73, that is, 80%). There is a relationship between the indicators of the problematic sides of banks and the reasons for the closure of the neighbouring inactive banks. For example, the area P includes active banks with low indicators of liquidity and capital adequacy, and inactive banks closed due to lack of capital and liquidity. We can also see that some inactive/closed

banks are located in the 'safe' area, below the R1 line. Most of them (9 out of 14 banks that are located below the line) are grouped near the border. The general causes of those banks recognized as insolvent are a lack of liquidity and capital adequacy. Thus, we can conclude that those institutions are attracted by the P area on the projection map but are held back below the R1 line as other indicators except liquidity and capital adequacy were mostly healthier than those for the banks in the P area (in the last period when they were active). The other 5 banks are located near A+, B+, and Corp areas left the market due to different reasons like consolidation with another bank, closing by the NBU due to low ownership structure transparency, or financial fraud. Such banks' actions generally do not have any visible effect on the indicators that were considered by us during this research, thus these banks were mapped in the 'safe' areas.

We can draw several important conclusions based on the results of only a visual analysis of a two-dimensional projection map obtained through t-SNE. First: the bank points located in the multidimensional space of the features we have selected really attract each other, having common features. This pattern is preserved when projecting data into a two-dimensional plane. Efficient and problem banks will "push off" from each other, locating on opposite sides of the centre of the map (in our case, these are the lower and lower left parts of the map, and the upper and upper right parts of the map), while mid-tier banks will be located near the centre of the map closer to investment-attractive banks. Second, problem banks do converge to fall into areas where closed banks were in the last period before they were liquidated. The location of a bank in an area of accumulation of already inactive banks is a worrying sign, and the reasons for the closure of the neighbouring inactive banks may indicate possible problem areas of active businesses. So, we can consider H2 hypothesis proved. Such an analysis will not be enough to prove hypothesis H1, thus we will further additionally use the clustering algorithm to test it.

The most adequate result when reducing the dimensionality was obtained with the following t-SNE settings: perplexity (regulate how to balance attention between local and global aspects of data) = 15, learning rate = 500, iterations = 1000, initialization with PCA, Barnes-Hut modification [42]. This configuration has been determined through numerous experiments with different hyper-parameters, and generally corresponds well with the recommendations of researchers [35]. At the same time, it can be argued about the sufficient ability of the algorithm to preserve both a local structure (banks with similar characteristics are located nearby in clusters) and a global data structure (clusters of investment-attractive and problem banks are significantly distant from each other). UMAP, in turn, has a greater focus on the global structure, and although individual clusters of points can be well separable, the points themselves within the clusters can be stuck together very close to each other..

Source: own study based on statistical data of the National Bank of Ukraine (NBU data, aggregated outstanding amounts on balance sheet accounts of the Ukrainian banks, 'Aggregation_eng.zip')

Figure 2. The results of reducing multidimensional data into a two-dimensional projection using UMAP (left box) and t-SNE (right box) for selected features for QI 2015 to QIV 2020

4.2. Clustering of Commercial Banks with Dimensionality Reduction Algorithms

In the previous approach, we used the t-SNE dimensionality reduction algorithm, and it showed good results for purely visual analysis of data for the last period. However, the results of this algorithm are poorly suited for further clustering (Figure 2, right box)

Figure 2 shows that the points on the low-dimensional projection form a field with a large number of groups of low-density points that are difficult to cluster regardless of the method chosen.

The UMAP algorithm, in turn, produces projections that are slightly less amenable to visual analysis, but it showed the highest quality result when used in conjunction with clustering algorithms. The following hyper-parameters were used for dimensionality reduction in UMAP: $n_neighbors$ (number of neighbours, controls the local vs global structure preservation) = 20, min_dist (minimum distance apart that points are allowed to be in low dimensional representation) = 0, $distance = correlation$. The correlation distance is calculated as follows:

$$1 - \frac{(u - \bar{u}) \cdot (v - \bar{v})}{\| (u - \bar{u}) \|_2 \| (v - \bar{v}) \|_2} \quad (3)$$

where u and v are two data arrays with two sets of coordinates for each data point; $\bar{u}$ and $\bar{v}$ are the mean of the elements of u and v accordingly; $(u - \bar{u}) \cdot (v - \bar{v})$ is the dot product of $(u - \bar{u})$ and $(v - \bar{v})$. We experimentally determined the optimal values of the parameters, relying on the official documentation of the UMAP library, which in turn is based on the work of the author of the algorithm [34].

Dimensionality reduction resulted in a map with projections of bank points onto a two-dimensional plane (Figure 2, left box). We used the Ward's Hierarchical Agglomerative algorithm. Only the setting of the number of clusters is required for clustering among the parameters. In order to determine the optimal number of clusters, we relied on the clustering quality metrics discussed in the Research Methodology section, and on the condition that the number of clusters should not exceed 30 (our assumption based on visual analysis of the two-dimensional UMAP projection in Figure 2). Having made runs and evaluating the quality of clustering for the number of clusters from 2 to 30, we found that the optimal number of clusters is 11. At the same time, the silhouette coefficient is 0.574, Davies-Bouldin score is 0.532, Calinski-Harabasz score is 6029.184, which can be generally considered high enough quality for clustering, given that we did not remove outliers from the data, but only smoothed them to preserve their impact on the division of banks into groups.

The table below describes the resulting clusters (Table

1). Based on indicators such as K0. Assets share in the system, K13. Credit losses provision ratio, K113. Bank business activity ration, K26. Dependence on natural deposits ratio, K31. Autonomy ratio (capital adequacy), K51. Highly liquid assets share, K68. Return on costs, we have formed our own simple ranking system and ranked banks by using it. The rankings attributed to the banks according to our system mostly coincide with the Minfin and the NBU rankings. The table for each cluster shows the average ranking for all banks in the cluster. Let us note that this ranking system does not claim to be completely objective and highly accurate. Although the rankings of banks are based solely on the ratios that are calculated on the basis of the current open statistics of the NBU, the ratios for calculating the rankings were selected subjectively, based on our vision of the importance of certain indicators for assessing the financial condition of banks. The purpose of the ranking system is not to assess banks as bad or good; it serves solely for the quantitative assessment of clusters formed in the course of simulation for their further comparison with each other. An important note here is that clusters presented in Table 1 are based on a mathematical approach, cluster analysis specifically. This approach differs from the one we used when analysing Figure 1 which was based fully on a visual analysis of the projection map. Thus, clusters formed in Figure 1 and clusters described in Table 1 are not much related to each other as they are using different approaches. However, both approaches provide fairly good outcomes in separating different classes of banks.

In Table 1, the features of the clusters, which we consider to be crater-forming based on the analysis of the ratios, are highlighted in bold. As we can see from the table, all banks can be conditionally divided into 3 large groups based on their financial condition and investment attractiveness. The first group is Class A (1st tier) banks (Cluster 1) or banks that can potentially be considered as such, but have minor problems in one or several indicators of their activity (Clusters 2-3). Such banks have the highest investment attractiveness and are reliable both for lending and for investing funds. The second group contains the largest number of banks (Clusters 4-6). Most of them can be classified as Class B (2nd tier) banks. They are divided into corporate and retail/universal (Clusters 4 and 6). The former may be of interest to clients looking for stable banks serving legal entities, and the latter may be considered attractive to the population as small and/or local banks. The rest of the banks (Cluster 5) are mid-tier banks that can hardly be called investment-attractive, but which are not unambiguously problematic. Those on the brink between the two categories can rather be characterized as risky. The third group is problem banks (Clusters 7-11). Each of the clusters has its own set of features that describe the banks that fell into it and indicate their problematic sides.

Table 1. Characteristics of clusters obtained from clustering of Ukrainian commercial banks for 2015-2020 using the Ward's Agglomerative Hierarchical Clustering algorithm.

#	N of banks	Characteristics	Average rating
1	5	Class A banks that have distinctive financial performance indicators.	8.00 / 9.00
2	5	Medium banks, potential Class A banks, most of which have problem assets in their portfolios.	6.20 / 9.00
3	1	Potential Class A banks, which, however, have problems with asset quality and capitalization.	6.00 / 9.00
4	10	Class B banks. Mostly medium-sized banks that are active in lending and investment activities, some of them have a problem with profitability.	5.50 / 9.00
5	12	Mostly medium-sized banks that do not stand out in terms of their performance. Most of the banks are pursuing a moderate or passive credit and investment policy and have problems with profitability. Some banks are undercapitalized and have problem assets in their portfolios.	5.08 / 9.00
6	10	Class B corporate banks. Mostly small banks, most of which are more focused on working with legal entities. Most banks have problems with profitability.	4.90 / 9.00
7	5	Systemically important banks, with a large number of assets, having problem assets in their portfolios and undercapitalization. Most of them have a low cost-benefit ratio.	2.80 / 9.00
8	7	Small banks, most of them focused on corporate lending, most of the banks conduct a moderate or passive lending and investment policy, all banks have problem assets in their portfolios and problems with profitability, 4 out of 7 banks are unprofitable.	2.57 / 9.00
9	2	Technical banks focused on specific goals within the public financial system. They practically do not conduct banking activities in the usual sense. They are low-liquid and unprofitable.	2.50 / 9.00
10	8	Mostly small banks with low liquidity, most of the banks conduct a passive lending and investment policy, have an unstable resource base, problem assets in their portfolios, and some banks are low-profit.	2.13 / 9.00
11	7	All banks pursue a passive policy, which is most likely justified by a small number of assets and a weak resource base. All banks have a significant share of problem assets on their balance sheets. 5 out of 7 banks are unprofitable, some are undercapitalized.	2.00 / 9.00

Source: own study based on statistical data of the National Bank of Ukraine (NBU data, Aggregated outstanding amounts on balance sheet accounts of the Ukrainian banks, 'Aggregation_eng.zip')

Having obtained sufficiently adequate results in the course of interpreting clusters, we can conclude that the separation of Ukrainian commercial banks (their clustering) based on a wide range of indicators of their activities is possible and efficient, investment-attractive banks are separable from problem banks, we proved our hypothesis H1.

It is still difficult to draw an unambiguous conclusion with regard to the issue of choosing a dimensionality reduction algorithm, even as a result of the research and tests carried out. However, we tend to reject hypothesis H3 that the UMAP algorithm is the most suitable for data dimensionality reduction in the study of Ukrainian commercial banks in terms of economic interpretation of the results. We justify this by the fact that the t-SNE algorithm, after optimizing its hyper-parameters, for visualization and subsequent economic interpretation of the resulting two-dimensional projection, fits even slightly better than UMAP, although UMAP was still preferable for cluster analysis.

5. Discussion

In our research, we relied on the works of researchers [19-21], which are closest to our study, focusing on dividing banks into clusters based on their financial performance indicators. However, they use small sets of indicators that describe mainly one side of the operation

of banks. In the work of researcher [20] the study is carried out in the context of the Ukrainian market, but it was carried out in 2016 and its results have lost their relevance, since the banking system has undergone changes since then. Besides, given that a system of bank separation, which would be based on a solid economic and mathematical methodology, has not been introduced and enshrined at the legislative level in Ukraine, we made an attempt to expand the studies presented above using the example of Ukrainian banks, involving a wider range of indicators, and propose an author's approach to dividing banks into groups/clusters and analysing their financial condition.

While selecting a general list of possible indicators, we analysed a number of Ukrainian- and Russian-language sources, but found a list of similar indicators in the works of researchers [19, 21, 43, 44, 45]. When selecting a basic set of indicators, the pairwise correlation matrix and the VIF metrics were one of the key indicators for us [31]. This approach is quite popular, and, for example, is used in the work of researcher [20].

When processing the data, we used the winsorization method and our experiments with its limits (tails) agree with the recommendations of researches [46-50]: a two-way limit of 1% for further work with the UMAP algorithm showed an excellent result. We also conducted a number of our own experiments with limits, reducing the dimensionality of different datasets, and found an improvement in the quality of t-SNE clustering at the

lower and upper limits of 2% and 5%, respectively. The use of dynamic indicators provides adaptive monitoring of the financial condition of banks [46]. This allows you to track changes in the dynamics and respond quickly to changes. An important role in this is played by central banks, which stimulate technology transfer [50]. Communication of central banks at the international level provides an exchange of new technologies and methods of assessing financial condition. This has a positive effect on the financial systems of the world and the international financial system as a whole.

In preparation for cluster analysis and data dimensionality reduction, we selected and adjusted the hyper-parameters of the algorithms and, as a result of numerous experiments, came to the same parameters that were proposed in the work of researches [35].

6. Conclusions

As a result of the study, we obtained a clustering of Ukrainian commercial banks, while successfully separating investment-attractive institutions from potentially problematic ones. We obtained 11 clusters, 3 of which we classified as those that include successful and efficient banks, and 5 more clusters we classified as those that include banks characterized by a risky level of performance. We obtained confirmation of our assumption that banks are located in the multidimensional space of diverse features describing their operation are next to the banks with similar features. It was also visually demonstrated that problem banks form a whole area on a two-dimensional map and are well discriminated against successful banks, as they are usually located on a two-dimensional projection in the diametrically opposite part of the map.

We carried out numerous experiments with the t-SNE and UMAP dimensionality reduction algorithms as part of the study. We came to the conclusion that the latter shows itself much better when paired with the hierarchical clustering algorithm. Provided that the parameters of both algorithms are optimized, t-SNE showed excellent results in a purely visual analysis of a two-dimensional map only for the last period. The results of our research may be of interest primarily to depositors (individuals and legal entities), Ukrainian and foreign investors. The approach and set of methods used in our work can be extrapolated to other countries, but one should be careful about the formation of a set of coefficients.

Of course, there are some limitations in our work. First, more than half of the coefficients were excluded from the initial set of features, because the open aggregated statistics of the NBU does not provide enough data to calculate them. The excluded features may be those that had a significant positive effect on the quality of reducing the clustering dimensionality. Second, we are dealing with unsupervised machine learning from the point of models, and have no actual data to compare the result. Assessing the quality of clustering with the metrics used can be subjective and is often applied rather to compare different clustering algorithms or their settings with each other. We relied on our simple ranking system to be able to compare our cluster scores with some actual value.

The use of other clustering algorithms to solve the problem of dividing banks into investment-attractive and problematic ones can be an interesting direction for further research. Another direction may be deepening this study through using detailed reports from each of the bank, thus calculating additional ratios. The application of the described complex of stages (pipeline) for the analysis of banking systems of other countries is also of scientific interest.

Appendix

Description of the coefficients used in the model (after winsorization and application of the robust scaler)

Name of the coefficient	Calculation	Brief description	Min	Max	Avg	Std
K0. Consistency ratio	Bank assets / Total assets in the system	Describes the consistency of the bank and its size	0.22	55.37	1.90	6.29
K12. Share of earning assets	Working assets / Bank assets total	Describes the quality of assets: the higher, the greater the efficiency of the use of resources and the business activity of the bank	-3.53	1.16	-0.19	0.86
K13. Credit losses provision ratio	Loan provisions / Credit exposure total	Banks tend to provide more reserves for low-quality assets. The higher the indicator, the potentially worse the quality of the loan portfolio	-0.94	15.34	0.92	2.81
K17. Return on assets ratio	Interest income / Income assets total	How many units of income are generated by one unit of working assets	-1.16	48.80	0.29	2.49
K19. Share of loans in assets	Total loans / Assets total	Describes the extent to which the bank is focused on lending in its active operations (receiving interest income)	-1.51	1.93	-0.06	0.7
K112. Share of securities in assets	Securities / Assets total	A high share of securities in the portfolio may indicate that the bank belongs to the category of investment banks and / or that there is free income	-1.48	4.22	0.28	0.86
K113. Credit-investment portfolio share	Credit-investment portfolio / Assets total	Describes the riskiness of the bank's credit policy	-3.32	1.17	-0.18	0.86
K24. Interbank lending rate	IBL / Deposits total	Describes the degree of dependence of the bank's resource base on interbank loans	-1.26	1.92	-0.08	0.72
K26. Private deposits dependency ratio	Private deposits / Liabilities total	Describes the bank's dependence on private deposits and the possible risks in case of panic of depositors	-0.45	8.62	0.87	1.75
K31. Autonomy ratio (capital adequacy)	Equity / Net assets total	Describes the extent to which the bank's risky investments are secured by its capital	-4.90	3.26	0.24	0.94
K32. Capital adequacy ratio	Equity / Deposits	Describes the degree of coverage of clients' funds by capital	-2.18	5.50	0.74	1.80
K33. Loan coverage ratio	Equity / Loans	Describes the ability to return the borrowed funds in case of non-repayment of loans	-2.49	5.97	0.50	1.55
K34. Return on capital (ROC)	Net profit / Equity	Describes the efficiency of using owners' funds	-12.36	44.86	0.17	5.03
K52. Highly liquid assets share	High liquid assets / Assets total	Describes the bank's ability to quickly pay claims of creditors and depositors	-6.05	4.04	-1.02	2.82
K65. Risk-free coverage ratio	Fee income / Costs total	Describes to what extent the risk-free income covers the total costs of the bank	-3.73	10.40	0.26	1.18
K68. Return on costs	Profit (before tax) / Costs total	Describes the cost efficiency of the bank, the amount of profit per unit of cost	-0.81	2.80	-0.03	0.64
K71. Tax burden ratio	Taxes total / Income total	Taxes per unit of income	-28.48	104.93	0.55	4.10

REFERENCES

- [1] Arrigoni S., Bobasu A., Venditti F., "The simpler the better: measuring financial conditions for monetary policy and financial stability," European Central Bank Working Paper Series, No 2451, 2020, <https://www.ecb.europa.eu/pub/pdf/scpwps/ecb.wp2451~c bf3f02232.en.pdf> (accessed Sept. 10, 2021).
- [2] Assaf A., Berger A., Roman R., Tsionas M., "Does efficiency help banks survive and thrive during financial crises?" Journal of Banking & Finance, vol.106, pp. 445-470, 2019. DOI: 10.1016/j.jbankfin.2019.07.013
- [3] Chen Y., Ting H., Wang M., "Government support and bank performance during the 2007–2008 financial crisis," The North American Journal of Economics and Finance, vol. 55, paper 101301, 2021. DOI: 10.1016/j.najef.2020.10 1301
- [4] Erfani G., Vasigh B., "The Impact of the Global Financial Crisis on Profitability of the Banking Industry: A Comparative Analysis," Economics, vol. 6, no. 4, paper 66, 2018. DOI: 10.3390/economics6040066

- [5] Rahman H. U., Pinto P., Hawaldar I. T., Tm R., Sarea A., "An Evaluation of Financial Performance of Commercial Banks," *International Journal of Applied Business and Economic Research*, vol. 15, no. 22, 2017, https://www.researchgate.net/publication/324750584_An_Evaluation_of_Financial_Performance_of_Commercial_Banks (accessed Sept. 10, 2021).
- [6] Stuhr, D. P. Van Wicklen R., Rating the financial condition of banks. Federal Reserve Bank of New York, New York, 1974.
- [7] Antoun R., Coskun A., Georgievski B., "Determinants of financial performance of banks in Central and Eastern Europe," *Business and Economic Horizons*, vol. 14, no.3, pp. 513-529, 2018. DOI: 10.15208/beh.2018.37
- [8] Zhavoronok A., Kholiavko N., "Banking System of Ukraine: Trends and Prospects of Development," *Modern Science*, vol. 10, pp. 129-142, 2020.
- [9] Afanasyeva O., Kammel A., "Ukrainian Banking Regulation: Its Challenges and Transition towards European Standards," *International Banking And Securities Law*, vol. 2, no. 1, pp. 1-81, 2017. DOI: 10.1163/24056936-12340005
- [10] Ramskyi A., Loiko V., Sobolieva-Tereshchenko O., Loiko D., Zharnikova V., "Integration of Ukraine into the European banking system: cleaning, rebooting and Basel III," *Banks And Bank Systems*, vol. 12, no.4, pp. 163-174, 2017. DOI: 10.21511/bbs.12(4-1).2017.05
- [11] Kamil I., Al-Mamory S., "Enhancement of OPTICS' time complexity by using fuzzy clusters," *Materials Today: Proceedings*, 2021. DOI: 10.1016/j.matpr.2021.06.441
- [12] Bakoben M., Bellotti T., Adams N., "Identification of credit risk based on cluster analysis of account behaviours," *Journal Of The Operational Research Society*, vol. 71, no. 5, pp. 775-783, 2019. DOI: 10.1080/01605682.2019.1582586
- [13] Gautam, R., "Determinants of Financial Performance: An Evidence from Nepalese Commercial Banks," *Amity Journal of Strategic Management*, vol. 1, no. 2, 2018, <https://amity.edu/UserFiles/aibs/39291.pdf> (accessed Sept. 10, 2021).
- [14] Zhou G., Sun Y., Luo S., Liao J., "Corporate social responsibility and bank financial performance in China: The moderating role of green credit," *Energy Economics*, vol. 97, paper 105190, 2021. DOI: 10.1016/j.eneco.2021.105190
- [15] Mihalovič M., "Performance Comparison of Multiple Discriminant Analysis and Logit Models in Bankruptcy Prediction," *Economics & Sociology*, vol. 9, no. 4, pp. 101-118, 2016, DOI: 10.14254/2071-789x.2016/9-4/6
- [16] Erdal H., Karahanoğlu İ., "Bagging ensemble models for bank profitability: An empirical research on Turkish development and investment banks," *Applied Soft Computing*, vol. 49, pp. 861-867, 2016. DOI: 10.1016/j.asoc.2016.09.010
- [17] Trung N. K. Q., "Determinants of bank performance in Vietnamese commercial banks: an application of the camels model," *Cogent Business and Management*, vol. 8, no. 1, 2021. DOI: 10.1080/23311975.2021.1979443
- [18] Rodriguez M., Comin C., Casanova D., Bruno O., Amancio D., Costa L., Rodrigues F., "Clustering algorithms: A comparative approach," *PLOS ONE*, vol. 14, no. 1, paper e0210236, 2019. DOI: 10.1371/journal.pone.0210236
- [19] Negnevitsky M., "Identification of failing banks using Clustering with self-organising neural networks," *Procedia Computer Science*, vol. 108, pp. 1327-1333, 2017. DOI: 10.1016/j.procs.2017.05.125
- [20] Kolodiziev O., Chmutova I., Biliaieva V., "Selecting a kind of financial innovation according to the level of a bank's financial soundness and its life cycle stage," *Banks And Bank Systems*, vol. 11, no. 4, pp. 40-49, 2016. DOI: 10.21511/bbs.11(4).2016.04
- [21] Dao B., Khanh P., "Cluster Analysis of Vietnamese Banks," *SSRN Electronic Journal*, 2014. DOI: 10.2139/ssrn.2543094
- [22] Bellman R. E., *Adaptive Control Processes*, Princeton University Press, Princeton, 1961.
- [23] van der Maaten L., Hinton G. E., "Visualizing High-Dimensional Data Using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579-2605, 2008.
- [24] Melit Devassy B., George S., Nussbaum P., "Unsupervised Clustering of Hyperspectral Paper Data Using t-SNE," *Journal Of Imaging*, vol. 6, no. 5, paper 29, 2020. DOI: 10.3390/jimaging6050029
- [25] Shubert E., Gertz M., "Intrinsic t-Stochastic neighbor embedding for visualization and outliers detection. A remedy against the curse of dimensionality?" *Heidelberg University*. 2017, <https://pdfs.semanticscholar.org/d767/2e61c99ee7d6c3711ca2b78469e9e4b17e43.pdf> (accessed Sept. 10, 2021).
- [26] Wattenberg M., Viégas F., Johnson I., "How to use t-SNE effectively," *Distill*, 2016, <http://distill.pub/2016/misread-t-sne> (accessed Sept. 10, 2021).
- [27] Vermeulen M., Smith K., Eremin K., Rayner G., Walton M., "Application of Uniform Manifold Approximation and Projection (UMAP) in spectral imaging of artworks," *Spectrochimica Acta Part A: Molecular And Biomolecular Spectroscopy*, vol. 252, paper 119547, 2021. DOI: 10.1016/j.saa.2021.119547
- [28] Becht E., McInnes L., Healy J., Dutertre C., Kwok I., Guanng L., Ginhoux F., Newell E., "Dimensionality reduction for visualizing single-cell data using UMAP," *Nature Biotechnology*, vol. 37, paper 38, 2019.
- [29] Yang Y., Sun H., Zhang Y., Zhang T., Gong J., Wei Y., Duan Y., Shu M., Yang Y., Wu D., Yu D., "Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data," *Cell Reports*, vol. 36, no. 4, paper 109442, 2021. DOI: 10.1016/j.celrep.2021.109442
- [30] Dixon W., "Simplified Estimation from Censored Normal Samples," *The Annals of Mathematical Statistics*, vol. 31, no. 2, pp. 385-391, 1960. DOI: 10.1214/aoms/1177705900
- [31] Dormann C., Elith J., Bacher S., Buchmann C., Carl G., Carré G., Marquéz G., Gruber B., Lafourcade B., Leitão P., Münkemüller T., McClean C., Osborne P., Reineking B., Schröder B., Skidmore A., Zurell D., Lautenbach S., "Collinearity: A review of methods to deal with it and a simulation study evaluating their performance," *Ecography*,

- vol. 36, no. 1, pp. 27-46, 2012. DOI: 10.1111/j.1600-0587.2012.07348.x
- [32] Ferreira A., Figueiredo M., "Efficient feature selection filters for high-dimensional data," *Pattern Recognition Letters*, vol. 33, no. 13, pp. 1794-1804, 2012. DOI: 10.1016/j.patrec.2012.05.019
- [33] Zhao Z., Liu H., "Spectral feature selection for supervised and unsupervised learning," *Proceedings of the 24th International Conference on Machine Learning - ICML '07*, 2007. DOI: 10.1145/1273496.1273641
- [34] McInnes L., Healy J., UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018, ArXiv, abs/1802.03426.
- [35] Kobak D., Berens P., "The art of using t-SNE for single-cell transcriptomics," *Nature Communications*, vol. 10, no. 1, 2019. DOI: 10.1038/s41467-019-13056-x
- [36] Oded M., Lior R. *Clustering methods*. Data Mining and Knowledge Discovery Handbook. Springer, Berlin, 2006.
- [37] Subasi A., *Practical Machine Learning for Data Analysis Using Python*, Elsevier, Berlin, 2020. DOI: 10.1016/b978-0-12-821379-7.00007-2
- [38] Ward J., "Hierarchical Grouping to Optimize an Objective Function," *Journal Of The American Statistical Association*, vol. 58, no. 301, pp. 236-244, 1963. DOI: 10.1080/01621459.1963.10500845
- [39] Rousseeuw P., "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53-65, 1987. DOI: 10.1016/0377-0427(87)90125-7
- [40] Calinski T., Harabasz J., "A dendrite method for cluster analysis," *Communications In Statistics - Theory And Methods*, vol. 3, no. 1, pp. 1-27, 1974. DOI: 10.1080/03610927408827101
- [41] Davies D., Bouldin D., "A Cluster Separation Measure," *IEEE Transactions On Pattern Analysis And Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224-227, 1979. DOI: 10.1109/tpami.1979.4766909
- [42] van der Maaten L., "Barnes-hut-sne," Researchgate, 2013, https://www.researchgate.net/publication/234119714_Barnes-hut-sne (accessed Sept. 10, 2021).
- [43] Albulescu C., "Banks' Profitability and Financial Soundness Indicators: A Macro-level Investigation in Emerging Countries," *Procedia Economics and Finance*, vol. 23, pp. 203-209, 2015. DOI: 10.1016/s2212-5671(15)00551-1
- [44] Dhiman R., "Identifying the key indicators of financial stability and financial development: a review of financial service sector," *Asian Journal of Management Science and Applications*, vol. 3, no. 4, paper 302, 2018. DOI: 10.1504/ajmsa.2018.098902
- [45] Kovalenko V., Kuznetsova L., Sergeeva E., Todorova N., Mylashko O., Tej J., "Cluster approach to banking supervision with reference to bank risk profile," *Economic Annals-XXI*, vol. 177, no. 5-6, pp. 82-91, 2019. DOI: 10.21003/ea.V177-07
- [46] Rudneva Y. R., Vanchukina L. I., Galeeva N. N., Gareeva Z. A., Rogacheva A. M., "Evaluation of Effectiveness of Commercial Bank's Activities by External Users with Application of Dynamic Indicators," *Espacios Revista*, vol. 39, no. 12, pp. 1-19, 2018.
- [47] Lv Q., Wang Y., Zhang D., Yang M., Zheng Y., "Performance Evaluation of Commercial Banks Based on Factor Analysis," *E3S Web Of Conferences*, vol. 214, paper 02017, 2020. DOI: 10.1051/e3sconf/202021402017
- [48] National Bank of Ukraine. 2021, <https://bank.gov.ua/en/statistic/supervision-statist/data-supervision#1> (accessed Sept. 10, 2021).
- [49] Steyerberg E., Vergouwe Y., "Towards better clinical prediction models: seven steps for development and an ABCD for validation," *European Heart Journal*, vol. 35, no. 29, pp. 1925-1931. DOI: 10.1093/eurheartj/ehu207
- [50] Gobareva Y. L., Gorodetskaya O. Yu., Karp M. V., Guliev I. A., "International Development Banks as Centers of Technological Exchange: Is Future Here?" *Universal Journal of Accounting and Finance*, vol. 9, no. 3, pp. 329-335. DOI: 10.13189/ujaf.2021.090306