Journal of Industrial and Systems Engineering
Vol. 16, No. 1, pp. 15-36
Winter (January) 2024

JISE

Reducing risk and increasing income of bank customers by Optimizing data mining for the multi-objective model of allocating facilities

Roya Cheshmikhani[1], Mohammad Ali Afshar Kazemi[2*], Abbas Tolouei Ashlaghi[2], Ezzatollah Asgharizadeh[3]

[1]Ph.D. Student in Industrial Management, Department of Industrial Management, Science and Research Branch, Islamic Azad University, Tehran, Iran.

[2]Full Professor, Department of Industrial Management, Science and Research Branch, Islamic Azad University, Tehran, Iran.

[3]Department of Management and economics , Tehran University, Tehran, Iran.

Abstract

Iranian Commercial banks are always considered as one of the most important institutions active in the money and capital market, due to the economic structure of the country and the lack of development of the capital markets, which makes them in charge of financing the economic sectors of the country. However, these banks are not successful in fulfilling their mission. There are several models such as linear programming, integer programming, and zero and one programming that can provide an optimal combination of the elements that make up the facility basket. However, entering financial information into mathematical planning by considering all conditions is not straightforward to achieve such a goal. In this research, using data mining to optimize the multi-objective model of facility allocation is done using neural network. In this article, First, the effective variables were extracted from the bank database and after preparation, the most important features were identified using different algorithms In order to cluster the customers, two clusters have been estimated and customers have been identified in two low-risk and high-risk categories .then low-risk customer divided to ten cluster. finally, by using convolutional neural network and combined this neural network by LSTM, the risk and profit of each cluster has been predicted. The accuracy of CNN-LSTM neural network was better. this method is better neural network for predicting profit and risk each of bank customers for payment of facilities. Combination CNN and LSTM in banking issues is innovation of this research.

* Corresponding Author
ISSN: 1735-8272,

Keywords: loan, data mining, clustering, deep learning, convolutional neural network, CNN-LSTM.

1- Introduction

The credit system is the cornerstone of any country's development. Granting facilities is an important part of every bank's operations, and this part of banking activities is economically important (Wang, 2018 , Nozari et al., 2012). In fact, economic growth and development are not possible without a quantitative increase of the "capital" factor as one of the factors of production, and for various reasons, it is not possible for all real or legal persons that in all cases and stages of their activity they can use their personal monetary resources to meet existing needs. Therefore, with their credit operations, banks provide the means to transfer resources from people who have directly invested in the bank to those who need money, and these people cause another group of people to be able to use these resources by repaying their installments. Failure to repay facilities on time causes the bank's resources to stagnate and in the long run causes the country's economic recession (G. Teles, J. J.P.C. Rodringues, R.A.L. Rabelo and S.A. Kozlov, 2020, Tavakkoli-Moghaddam et al., 2022).

Commercial banks of Iran have always been among the most active institutions in the money and capital market and are responsible of financing the economic sectors of the country. Currently, commercial banks in Iran are not very successful due to the unprincipled allocation of depositors' resources to facility recipients. It is obvious that if a model is not designed for credit portfolio management and in other words, if the allocation of facilities granted is not done properly in the portfolio, the bank will face bankruptcy in the long run (Bernanke, 2009)); (Armantier, 2015, Gharachorloo et al., 2021, Nozari, 2024). Because it allocates a large budget and capital to some of its facilities, which will increase credit risk and consequently the probability of non-recovery of claims. The issues and problems of facility portfolio management have been the most important reason for bankruptcy or loss of banks and financial and credit institutions.

Most of the existing methods related to the four models proposed as credit portfolio models are divided into three categories: *structural models*, *models based on macroeconomics*, and *models based on past performance*. Structural models use data and items of financial statements and changes in the value of assets in the capital market. Models based on macroeconomics calculate credit risk based on the correlation and relationship of industries with some macroeconomic variables. Statistical models, which are based on past performance, have measured credit risk in banks using some statistical distributions. Therefore, the consideration of profit and risk of each of the bank's customers in the case of granting facilities has not been taken into account in any of the previous models and methods. In this research, it can be done by trying to predict the behavior of banks regarding the re-creation of payments, firstly by feature selection the customers of bank are clustered , then the amount of risk and income of each customer is obtained by using the integrated neural network algorithm. CNN-LSTM never has been discussed for predicting behavior customers of banks, and this is one of the innovations of this research, which will discuss in the following.

2- Literature Review

Considering that identifying customers is one of the most important steps for any organization to provide services so that the organization does not incur excessive costs, many researches have been conducted in this regard as described below.

Zhongying Yang and Xiaolong Su (2012) have provided better services to customers by deep analysis on customer behavior. To this end, using Support Vector Machine (SVM), they clustered customers and analyzed the services and products used in each cluster. However, they do not provide any solution to improve or employ this scheme.

Skitavsky Sabo (2014) has tried to find a solution for the points that are located on the borders of the clusters in addition to clustering different data sets. That is, a good sensitivity analysis has been done. For this purpose, he has proposed an efficient strategy for analyzing and investigating these points. However, a major drawback of this method, which is based on k-means, is the uncertainty about the optimality of the obtained answers due to the random selection of the initial cluster centers, which happens when the data dispersion is high.

Tatnia Hubnowa (2014) has investigated the strategies of granting facilities to non-governmental sectors and the problems and challenges related to the policy of granting facilities to these sectors. In this article, the credit risk of two Bulgarian banks in 2014 has been studied. This research shows that the risk in the money and capital market is different and to reduce the credit risk, the return on investment in these banks should be increased. It means that risk and return have opposite aspects and in order to reduce credit risk, the investment and return of the investor in the monetary sector must be increased.

By developing a standard method and with the help of the C-Means algorithm, Indrani Bass (2015) clustered the customers who are dissatisfied with the mobile bank services and then analyzed their behavior at specific times to leave the system using the available data pattern and specifies the reasons for it.

Yasmin Bashir Malim (2017) has investigated the impact of risk management on banking operations in Islamic banks in Kenya. In this research, attention has been paid to the operational risk of banks in conditions of uncertainty. In granting facilities, the rules of Islamic banks are very clear and the variables considered for the granting facilities are divided into two categories, controllable and non-controllable. The profitability of banks has been examined based on these balanced variables and it shows that with a 95% confidence factor, the increase in the profitability of Islamic banks in Kenya is a result of reducing the operational risk and the optimal combination of variables in granting facilities. Finally, the conducted studies have shown that Sukuk facilities are the best type of facilities in Islamic banking.

Irina Yanenkova et al. (2021) addressed the issue of bank credit risk management using the cost risk model. Bank credit risk modeling was proposed based on neural network technologies that provide high reliability in determining credit risk. This research aims to improve and develop methodical support and scientific recommendations for reducing the level of risk with the value-at-risk method and combining it with fuzzy planning methods.

Azarnoosh Ansari (2016) has used a combination of c-means fuzzy clustering and genetic algorithm to cluster steel industry customers. In this research, customers were divided into two clusters using LRFM model variables (length, speed, frequency, and monetary value). The results of this article showed that the customers belonging to the first cluster had more long-term relationships, business efficiency, and business frequency, but the monetary value was lower than the average values of this criterion for all customers. The results also showed that the customers of the second cluster were commercially higher than the commercial and monetary value, but the length of the relationship and the frequency of the business were lower than the average values of these criteria for all customers. It was also found that the combined algorithm (i.e. c-means fuzzy clustering and genetic algorithm) used in this research had a linear Mean Square Error (MSE) compared to c-means fuzzy clustering. It is unclear how the number of clusters is chosen and the reason for using genetic algorithm is not justified in this work.

Mehdi Khosroviani et al. (2022) predicted the liquidity risk of Iran's private banks using an artificial neural network model. To this end, using a multi-layer perceptron neural network, they have predicted the liquidity risk of private banks from 2018 to 2020 using MATLAB software (Khosroyani, M. 2022).

By reviewing recent research in the field of banking and optimal combination of facilities payment, it is clear that most research has focused on reducing operational risk to increase profitability and optimal combination of investment portfolio in the capital market and less attention has been paid to the discussion

and examination of optimal combination in the money market. In addition, most of the researches that have used operations research and statistical methods have been about industrial issues and less use has been made of operations research in financial and investment issues. In addition, in the optimal portfolio of facilities or investment combinations, mostly genetic algorithm has been used and less use has been made of other fuzzy methods in uncertain conditions.

In addition, in Table 3, we can mention some of the important studies under study, which in the description section of the table, a summary of them is mentioned along with the similarities and differences with this research, and their limitations and strengths are mentioned in the order of the year of publication.

Table 1: Comparison of some previous research with the current research

Row	Title	Researchers	year of publication	Journal
1	Innovative Application of Artificial Intelligence Technology in Bank Credit Risk Management	Shuochen Bi, Wenqing Bao	2024	International Journal of Global Economics and Management
	In this article, it is stated that in recent years in Silicon Bank, the ratio of loans has decreased year by year, while the ratio of bonds has increased year by year. This change in asset structure makes banks more dependent on the performance of the bond market, which is highly volatile. In addition, Silicon Bank relies too much on low-interest deposits in the market to attract deposits. These types of deposits may be quickly drained as soon as market liquidity is tight. Risk for banks At the same time, with the increase in the interest rate of the Federal Reserve, the demand of depositors for interest income is also increasing, which has led to an increase in the cost of Silicon Bank's debt. This incident also serves as a reminder to other banks to take the changes in the market environment very seriously during their operations. Allocate assets and liabilities rationally and reduce risks. This article deals with bank credit risk management through deep learning and big data analysis and artificial intelligence and is similar to the upcoming research.			
2	Digital Bank Runs: A Deep Neural Network Approach	Marc Sanchez-Roger , Esther Puyol-Antón	2021	Sustainability, Digital Transformation and Fintech: The New Challenges of the Banking Industry
	This paper has developed a deep neural network (DNN) design to assess the potential impact of the introduction of CBDC on the banking sector, in particular, it focuses on the link between CBDC and bank management phenomena. This work presents an innovative method to assess the consequences of introducing a CBDC, which allows simulating the transfer of wealth between different financial assets depending on the design of a CBDC. In this research, the neural network method is used, with the difference that the neural network is DNN, which is the opposite of the current research, which uses hybrid neural network models such as CNN and CNN-LSTM, but the data of this article is for digital currency, which is different from the data of the current research.			

3	Development of Mathematical Models for Predicting Customers Satisfaction in the Banking System with a Queuing Model Using Regression Method	Abiodun	2017	merican Journal of Operations Management and Information Systems
	This article calculates the satisfaction and performance of customers in the bank's queuing system by using three-day data of customers in First Bank of Nigeria PLC and predicts their satisfaction by using linear and non-linear mathematical models. In determining and defining the indicators of customer satisfaction in this research, this article has influenced the proper and different view of how to measure customer satisfaction by presenting data in such a way that it has predicted the level of customer satisfaction without using a questionnaire and using mathematical models. The similarity of this model with the research It is present in multi-objective linear mathematical modeling, but its main difference is in applying the level of customer satisfaction and allocating bank services and products considering its increase.			
4	Analysing customer behaviour in mobile app usage", Industrial Management & Data Systems	Chen,Q. ,Zhang, M. , Zhao,X	2017	Industrial Management & Data Systems
	In the article by Chen et al., it shows the application of the RFM method in customers who use mobile phone applications in a way that identifies the customers' behavior in order to provide the required services and programs and also prevents the occurrence of crimes in the use of financial services. The rules used in this article to prevent suspicious financial transactions were significant and one of the applications of the RFM method in data mining has been well described. In this research, other data mining methods have been used.			
5	Detecting the migration of mobile service customers using fuzzy clustering	Indranil Bose, Xi Chen	2015	Information & Management
	This article, by developing a standard method and with the help of C-Means algorithm, deals with the clustering of customers who are dissatisfied with the mobile bank services, and then, using the available data model, analyzes their behavior at specific times to leave the system and identifies the reasons for it. that the researcher has benefited from its clustering model in this research			

In most of these researches, in the application of clustering, there is more qualitative analysis and lack of practical use in decision-making problems, and feature selection has not been used in clustering, while this gap has been resolved in the current research. The most important innovation of the current research is the use of integrated neural networks such as CNN-LSTM in banking topics, which has not been done in previous researches.

3-Metodology

The purpose of this research is to provide a model for proper payment of facilities to each group of bank customers. For the proposed model, real data was used, which was collected in the field from branches of Shahr Bank in Iran. Also, the variables used in this research are widely used in scientific methods, and it is possible to access the clean data available for each of them in the bank databases. The variables were selected from Shahr Bank's database. One of the most important reasons for using these variables is the comprehensiveness of each variable because they fully cover the research objectives. In this research, the characteristics of bank customers such as gender, income, and occupation, and the variables affecting the facility, such as the interest rate, the duration of the facility, and the loan amount, have been used. Also, many of the customer's basic and personal information such as education, geographic location, occupational background, capital level, e-mail, and telephone number cannot be used and have been deleted. This is because of the scarcity and low quality of this information in the database despite their possible indirect and imperceptible effect on the research objectives.

The format of the dataset is in the form of a CSV file and data analysis was done using R software and Python programming language. The data used in this research has 19 features and 1650 records. Among these variables, some of them are missing values. The problem of missing values in data science and especially data mining occurs when one or more observations have unrecorded or missing values in the "data frame" columns. In this case, we say that the observation has a missing value or a missing value (Bhandari, 2022). To perform statistical analysis on datasets with missing data, we need to determine the role of such observations in the calculations related to statistical analysis. To introduce missing data in a vector, we use the expression NA. To check the missing data in a vector, the function na.is() is used. This function provides a logical vector to the length of the desired vector, with the missing values corresponding to the members of the vector indicated by TRUE (Howell, 2021). Figure 1 shows the percentage of missing values for each variable.

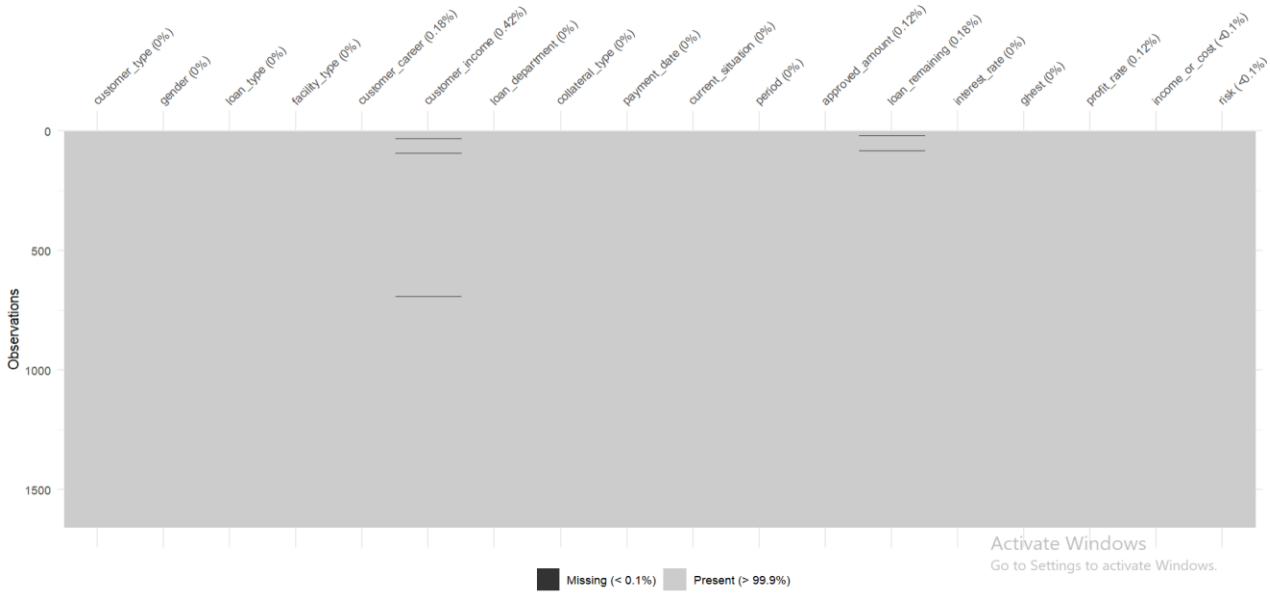

Figure 1: Missing Values

According to this figure, the percentage of missing values for the mentioned variables is as described in Table 2.

Table 2: Percentage of missing values

Row	Variable name	Persentage of missing value
1	Customer income	0.42%
2	Customer career	0.18%
3	Loan remining	0.18%
4	Approved amount	0.12%
5	Profit rate	0.12%
6	Income or cost	<0.1%
7	Risk	<0.1%

3-1- Placement Missing Value by The k-nearest neighbor algorithm

There are several ways to deal with missing data, and removing instances of missing data is perhaps one of the easiest. However, in this method, the number of samples is reduced. This is while in data mining, the more samples, the more accurate the modeling is. Sometimes it is possible to value the empty values using

different algorithms. In this research, the k-nearest neighbor method is used. The k-nearest neighbor algorithm is a non-parametric statistical method used for statistical classification and regression (Huang, 2018). In the classification mode, according to the specified value for k, the distance of the point we want to label with the nearest points is calculated(Guo et al, 2003), and according to the maximum number of votes of these neighboring points, we make a decision regarding the label of the desired point(Wang, 2022). Different methods can be used to calculate this distance, where the Euclidean distance is one of the most prominent of these methods (Ray, 2019). In the regression mode, the average value obtained from k is its output (Zhang, 2016). Since the calculations of this algorithm are based on distance, data normalization can help improve its performance (Du and Li, 2019). In this research, using the k-nearest neighbor, the average value of each column has been replaced instead of the missing values. Figure 2 shows that missing values have been removed.

Figure 2: Missing values

The input of most machine learning models such as neural networks must be numeric. For this purpose, qualitative data must be converted to numeric in the data preprocessing stage. In machine learning models, for data that is categorical (for example, a dataset of customer type attributes), we must convert them to numeric before applying the model to them. Labeling is one of the methods of converting qualitative data into quantitative data. In this research, qualitative features have been converted into quantitative data using this method.

3-2- Data Normalization

In the next step, the normalization of the data is done. Data normalization is a method for uniforming the range of values related to different variables of the research and is also known as data scaling. If the unit of measurement of the studied variables is diverse, the data can be scaled using normalization methods. Normalization or de-scaling is an underlying concept in multi-criteria decision-making methods such as AHP and ANP, and it enables the comparison of data with different measurement criteria (Henderi, 2021). Another concept of normalization, which is also known as standardization, is used in artificial neural network analysis and data envelopment analysis. In this research, normalization has been calculated using the z-score function, which is calculated based on the distance of the data from the average and standard of

the data (Zach, 2021). To use this method, the average and standard of the dataset must be calculated first. Then, for each data, its distance from the mean is calculated using formula (1):

$$z = (x - \mu) / \sigma \quad (1)$$

Where, x is the desired data, μ is the average of the data, and σ is the standard deviation of the data. After normalization, the distribution of each dependent variable will be according to Figure 3.

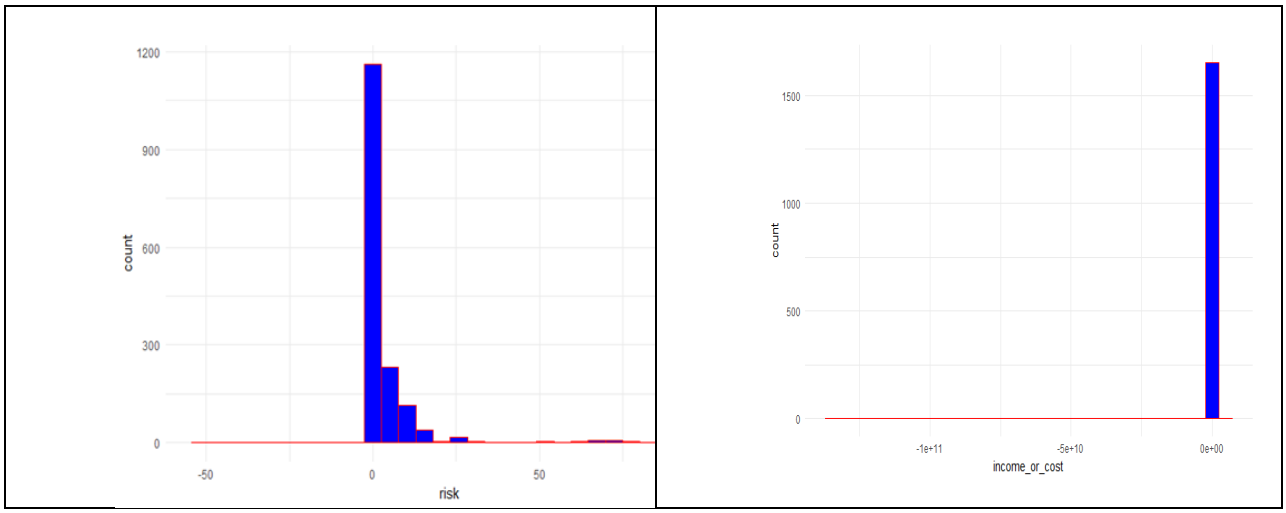

Figure 3: Distribution of dependent variables

The diagram in Figure 3-a is the normalization of the income-cost variable and the diagram in Figure 3-b is also related to the risk variable after normalization. As it is known, after normalization, the graph has a skewness.

3-3- Feature Selection

After preparing the data, the most important features that are effective in repaying the facilities have been investigated. Feature selection methods have become an inseparable component of the learning process when faced with high-dimensional data. A correct feature selection can lead to the improvement of the inductive learner from various directions, including learning speed, generalization capacity, and simplicity of the inferred model. To deal with the problem of a high number of features, dimension reduction methods are necessary and can help to improve learning efficiency.In this research, three methods have been used to select features.

3-3-1- Random Forest Algorithm

The first method is to use the random forest algorithm. As the name of this algorithm suggests, the random forest algorithm is a classifier that includes a number of decision trees in different subsets of the dataset and is averaged to improve the prediction accuracy of that dataset (Hajjem, 2014). Instead of relying on a decision tree, the random forest makes predictions from each tree based on the majority of votes (Nagifor, 2019) and considers the final result as the output. The more number of trees in the forest leads to higher accuracy and avoids the problem of overfitting (Speiser, 2019).After finding the best number for the trees, the selection of features for the two dependent variables of income-cost and risk is described in Figure 4.

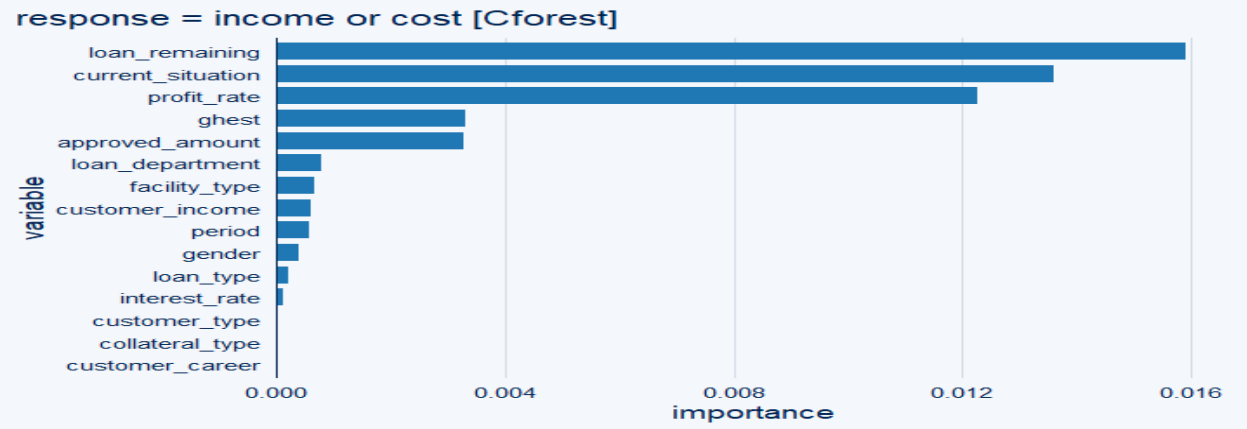

Figure 4: Random Forest Algorithm for Income/Cost

As shown in Figure 4, the effective variables on the dependent variable in the random tree have been identified in order from the most effective to the least important. The balance of the facility amount, the interest rate of the facility, the repayment time of the facility, the amount of the facility, the income of the customer, the installment of each of the facilities, the rate of the facility, the type of loans paid, and the type of the customer are among the most effective variables for the income-cost dependent variable.

Figure 5 shows that variables such as the remainder of the facility amount, the current situation, the facility interest rate, installments of each of the facilities, the facility amount, the type of collateral, customer income, the duration of facility repayment, gender, type of loans paid, and interest rate affect the risk variable.

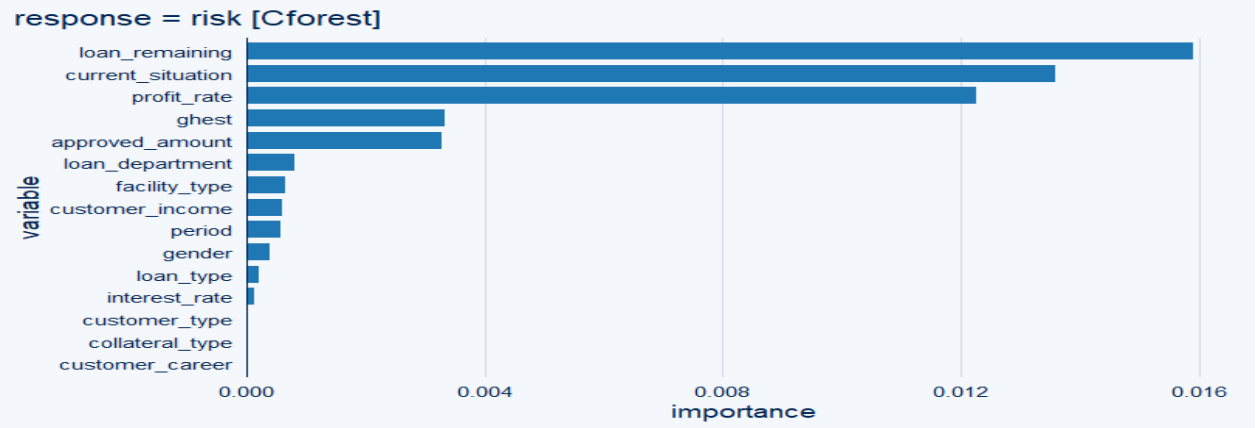

Figure 5: Random Forest Algorithm for Risk

3-3-2- Multivariate Adaptive Regression Splines (MARS) algorithm

The second method used in this research to find the most effective features is the use of the Multivariate Adaptive Regression Splines (MARS) algorithm. Despite the complexity of the MARS algorithm, the main idea is relatively simple, which is to replace discontinuous branches with continuous passes modeled by a pair of straight lines. At the end of the model-building process, the straight lines in each node are replaced by a very smooth function called *spline*, which results in the new divisions not dependent on the old ones (De Andrés et al, 2011).

Unfortunately, this means that the MARS algorithm does not have the CART tree structure and cannot create rules. On the other hand, MARS automatically finds the most important predictor variables and also the interaction between them and then determines the dependency between the response and each predictor. The result of the automatic regression tool is automatic and step-wise (Adnan, 2020). MARS, like most neural networks and decision tree algorithms, tends to overfit the training data (Dormann, 2013). This problem can be solved in two ways. First, cross-validation is done manually and the algorithm is adjusted to produce good predictions on the test set. Second, there are different adjustment parameters in the algorithm itself that guide the internal cross-validation. Using the MARS algorithm in this research, the variables that have the greatest impact on the output are according to Figure 6.

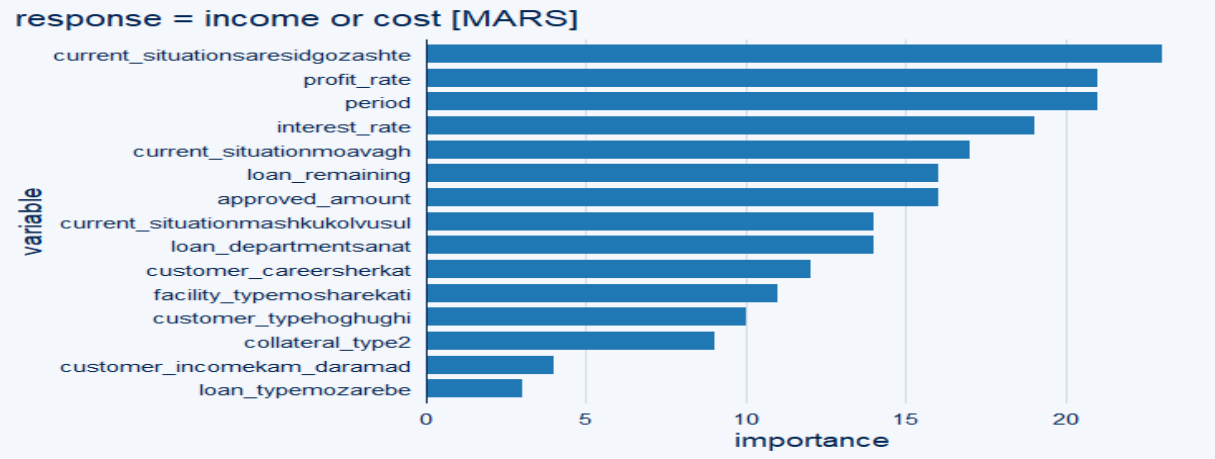

Figure 6: Income/Expense of MARS Algorithm

In this algorithm, it was determined that the influencing variables on the income-cost variable, in order of priority, are: the current status of the facility (overdue), the interest rate of the facility, the current status of the facility (arrears), the sector of the facility (industry), the customer's job (company), type of facility (participatory), customer type (legal), collateral type (second type), low-income customers, and facility type (Mudarabah).

According to the MARS algorithm, as depicted in Figure 7, the variables affecting the risk of the facility payment are as follows: the amount of the facility balance, the current status of the facility (overdue), the interest rate of the facility, the duration of the facility payment, the current status of the facility (arrears), the facility section (housing), the interest rate of each facility, the current status of the facility (doubtful availability), the amount of the facility paid, the type of collateral (the second type), customers with low income, and customers with medium income.

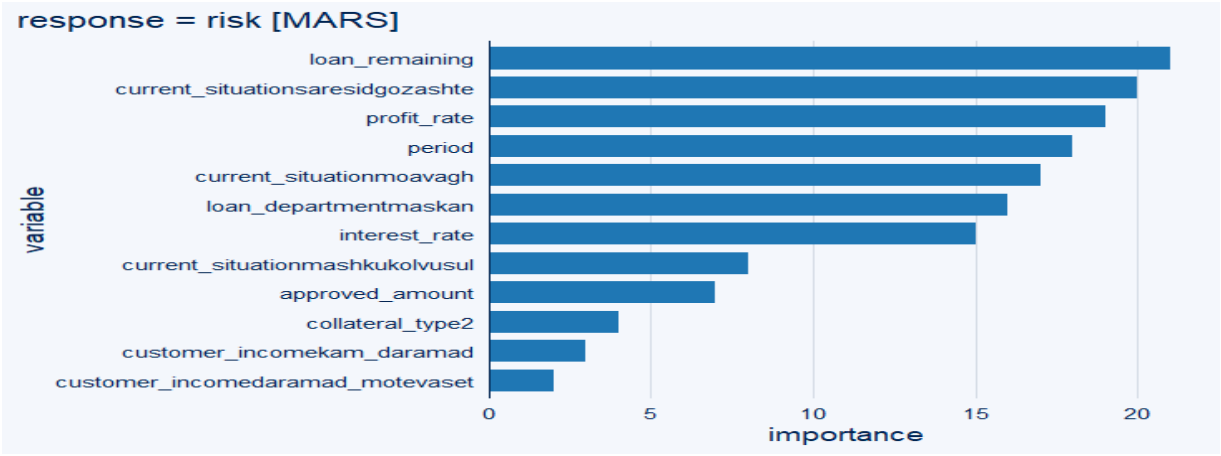

Figure 7: Risk in MARS Algorithm

3-3-3- the stepwise multiple regression algorithm

The third algorithm used in this research to select features is the stepwise multiple regression algorithm (Miller, 2002). In this model, predictor variables are added to the equation one by one, and then removed if they do not have a significant role in the regression. In the forward selection method, predictor variables are entered into the equation one by one if they meet the entry criteria and are not removed after entering. In the backward selection method, all predictor variables are first entered into the equation and then removed one by one if they do not have the criterion required to remain in the model. The step-wise method is a combination of the previous two methods and is recommended as the best method. As can be seen in the step-by-step regression method (Figure 8), some variables have become negative due to the inverse effect they have on the income-cost response variable.

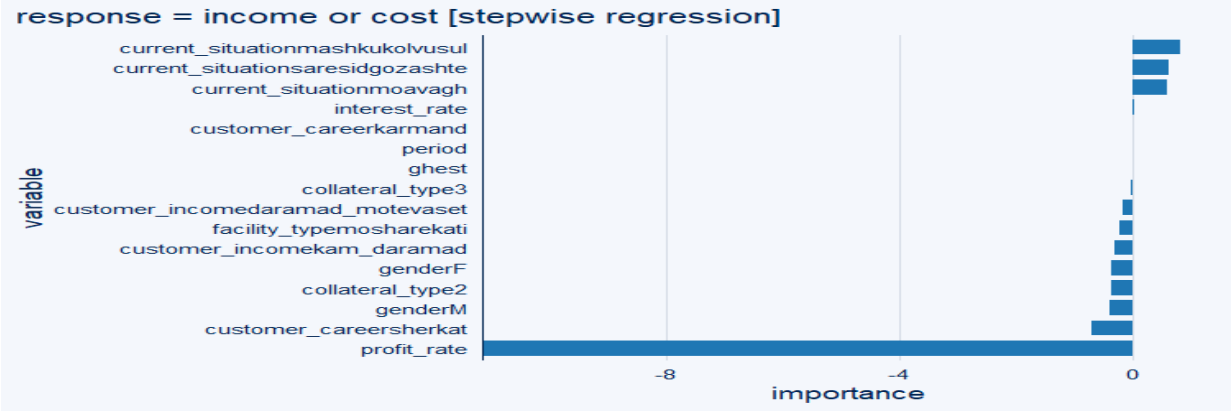

Figure 8: Income/Cost in Stepwise Regression Algorithm

Figure 9 shows the variables that have a direct effect on the risk variable in the order of influence, which are: the current status of the facility (respectively: overdue, past due, doubtful) and the type of facility (Mudarebah).

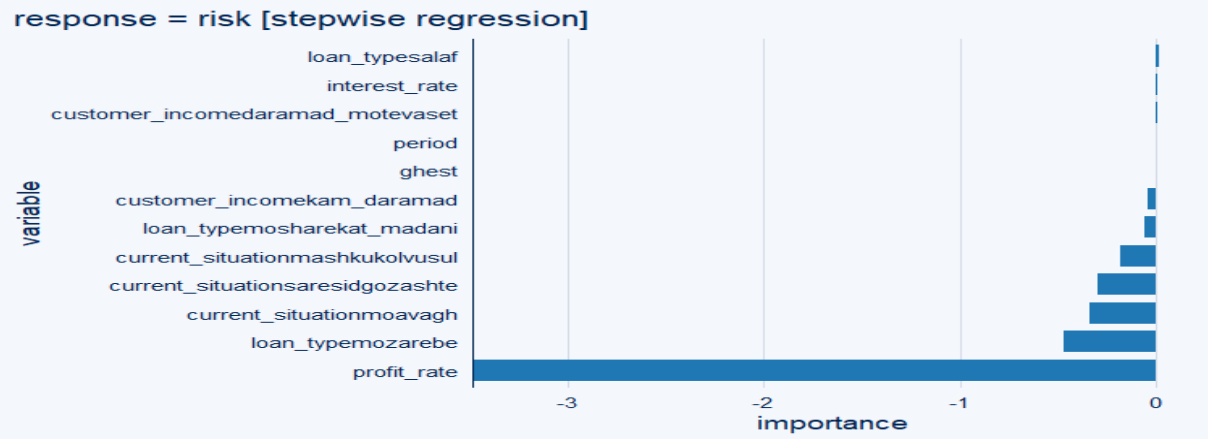

Figure 9: Risk in Stepwise Regression Algorithm

3-4- clustering

In this research, bank customers were first clustered with the characteristics that were provided to us from the bank's data. Clustering is a process that allows you to divide a set of objects into separate groups. Each division is called a cluster. The members of each cluster are very similar to each other according to their characteristics, and instead, the degree of similarity between the clusters is minimal. In such a case, the purpose of clustering is to assign labels to objects that indicate the membership of each object to the cluster. The main difference between clustering and classification is that there are no initial labels for the observations. In classification methods, there is a label for each sample, and classification and categorization can be done based on these labels. But in clustering methods, these labels are not present, and the separation criteria will only be the degree of similarity of each of the samples. Therefore, clustering is considered one of the unsupervised machine learning methods (Chen, 2017) and classification is a supervised learning method.

3-4-1- optimal number of clusters

In this research, using the silhouette criterion, the optimal number of clusters in R-Studio software has been obtained as 2 clusters. The silhouette method is one of the most common and best validation methods of the k-means clustering algorithm, which was first proposed by Kaufman and Rocio in 1990.

In this formula, the correlation a(x) is the average distance (x) to other vectors in the same cluster, the separation index b(x) shows the average distance (x) to other clusters, and S(x) is a measure of how close each point is in a cluster, which is scored relative to its neighboring clusters. The scoring range of this function is between 0 and +1. In this way, if this measurement is in the range of +1, it indicates that the cluster in question is far away from its neighboring cluster, and the zero state indicates that there is no separation between the cluster in question and the neighboring clusters. Finally, an output of -1 indicates the probability of wrongly assigning the desired cluster. The result obtained from the above method is as follows:

between_SS / total_SS = 75.4% (2)

Acording silhouette method that is shown in Figure 10, the best number of clusters is two.

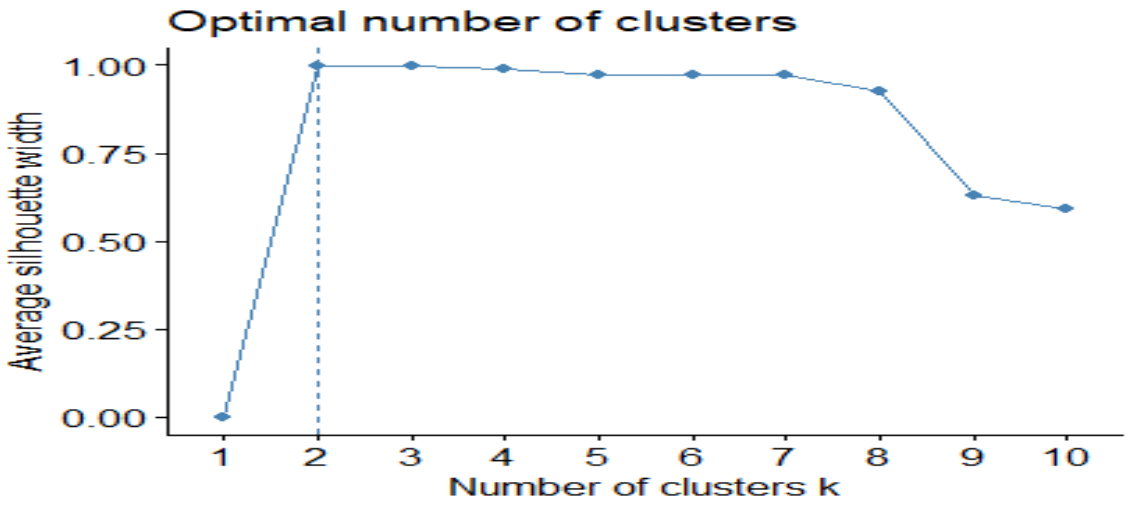

Figure 10: Optimal number of clusters

In this research, two k-means and K-Medoids algorithms are used for clustering.

3-4-2- K-Means clustering algorithm

The k-means clustering algorithm is an unsupervised learning algorithm (Mohammadi and Zanganeh, 2016). The k-means clustering algorithm is considered from the group of discriminative clustering methods and its computational complexity is equal to $O(n^{dk+1})$, provided that n is the number of objects, d is the dimension of features, and k is the number of clusters. Also, the time complexity for this algorithm is equal to O(nkdi), which, of course, means the number of iterations of the algorithm to reach the optimal solution (Bock, 2007). Here, the variables mentioned earlier in the selection of features are used for clustering. As it is known in the dplyr function, variables of payment facility, facility balance, installments, facility interest rate, cost or income, and risk have been used. The number of clusters is considered to be 2 and nstart must be a number above 15. There are 3 algorithms in k-means, which are: Lloyd's algorithm, Forgy's algorithm, and Hartigan-Wong's algorithm. All 3 algorithms have been used, and the output of k-means is shown in Figure 11.

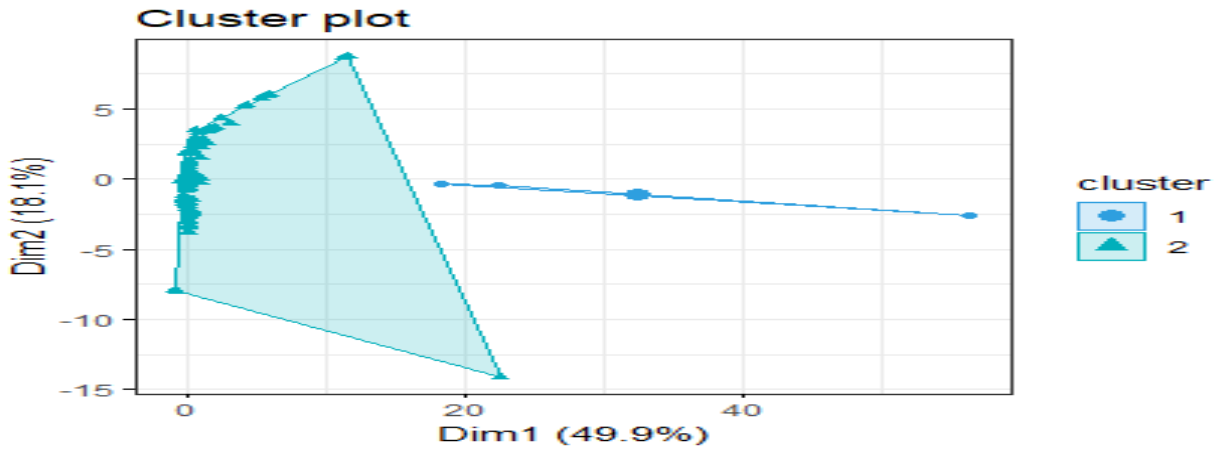

Figure 11: K-Mean clustring

3-4-3- K-Medoids clustering algorithm

In the next step, the k-medoids algorithm was used for clustering. The k-medoids algorithm is an object-based algorithm that selects representative clusters from the data itself and not averaging them. In fact, the middle of a cluster is the most central element of a cluster (Zhang, 2005). The purpose of this method is to reduce sensitivity to large values in the dataset. In this algorithm, each cluster is introduced with one of the data close to the center. As seen in Figure 12, the optimal number of clusters is 2, regardless of using Euclidean (Estivill-Castro, 2001) or Manhattan distance for clustering.

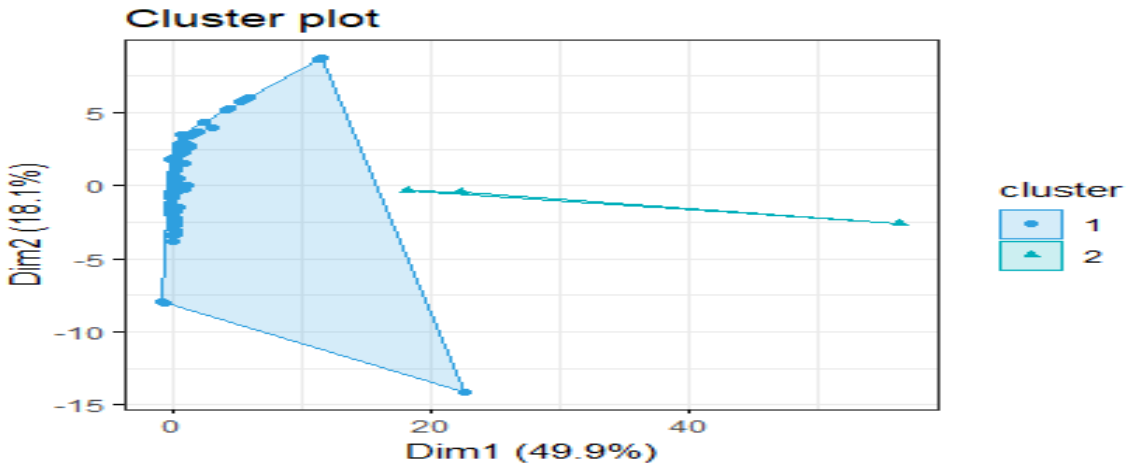

Figure 12: K-Medoids clustering

The result of clustering using the k-medoids method is as follows:

```
Medoids:
       ID approved_amount loan_remaining       ghest profit_rate income_or_cost     risk
[1,] 1148    5.000000e+08   4.157526e+08    17889101  0.09369886        7893890  0.00000
[2,] 1424    1.204945e+12   1.600629e+12 93028258643  0.06284257   -24865811247 16.41918
```

These results indicate that customers have been classified into two groups of good customers and bad customers, who have positive income and zero risk and bad customers have a risk of 16.42 and negative income or in other words have a cost for the bank. Obviously bank decided to pay loan just for good creditworthy customers.

3-5- Clustering of Low-Risk customers

In this research the creditworthy customers were classified into ten clusters by the Elbow criterion. In this method, the incremental values of k are drawn on the horizontal axis and the sum of errors that occurred when using the average k is drawn on the vertical axis. The purpose of using this method is to find a k that does not increase the variance too much for each cluster. Figure 13 shows the optimal number of clusters, which are 10.

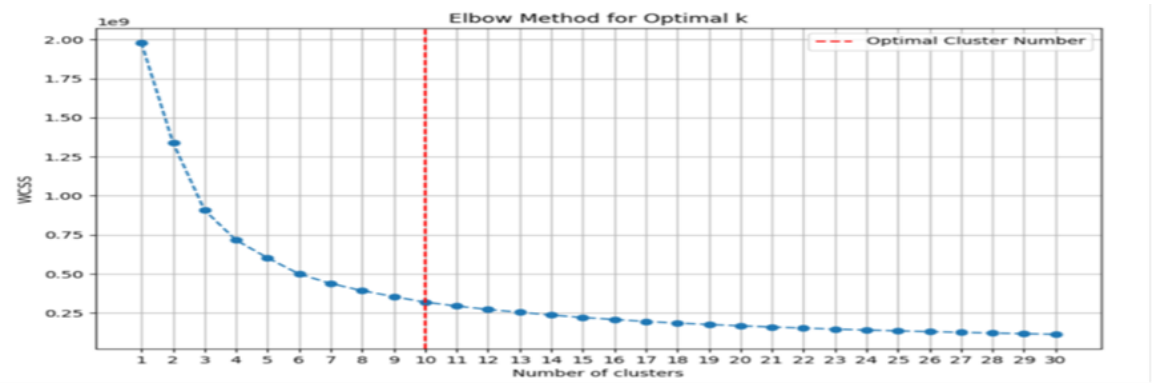

Figure 13: Optimal number of clusters foe good customer

According to the K-Means method, customers are divided into ten clusters as shown in the figure 14.

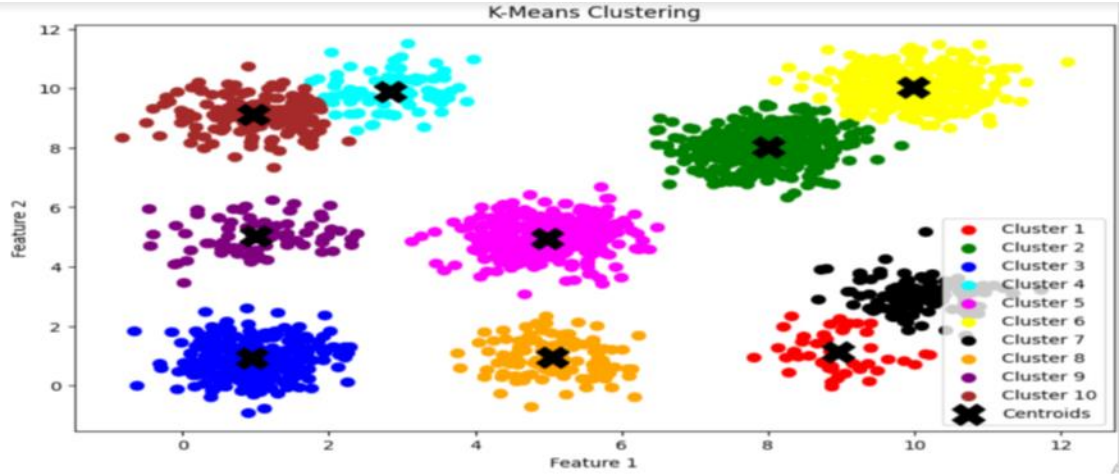

Figure 13: K-Mean clustring

3-6- convolutional neural network[†]

Finally, we predict the behavior of customers using convolutional neural network (Dormann, 2013) or CNN. Convolutional neural network is a very efficient method in deep learning, which is effective in predicting and dividing weights (Al Machot, 2019). For this purpose, first, the required libraries for designing the convolutional neural network model were called in Python software and then the data set was divided into training, testing, and validation. In general, the training part is used to train machine learning models, the testing part is used to evaluate the performance of the models on new data, and the validation part is used to adjust the model parameters and select the best model among several different models (Hjelkrem, 2023). The random method is usually used to divide the dataset into three parts (Zhou, 2022). In this project, 70% of the data is allocated to the training part, 15% to the testing part, and 15% to the validation part. The model is defined layer by layer. At first, a convolution layer with 64 filters, kernel size 3, and ReLU activation function is defined. Then, a MaxPooling layer with a size of 2 is defined, which is used to reduce the dimensions of the data. This layer halves the dimensions of the data by selecting the largest value in both input data values. To avoid overfitting, a Dropout layer with a value of 0.5 is defined. This layer is used to randomly remove the learning weights of the model, and here the value of 0.5 means

1-CNN

that in each period 50% of the model weights are randomly removed. After that, a Flatten layer is defined, which is used to transform the input data into one-dimensional data. This layer linearly transforms the input data into one dimension so that it can be fed to the next Dense layer. Finally, two dense layers are defined. The first layer is defined with 128 neurons and relu activation function, and the second layer is defined with two neurons and no activation.

To compile the model, three optimizer, loss and metric parameters are used for the convolutional neural network model. The optimizer parameter is the optimization algorithm that is used in the process of training the models. Here, Adam's optimization algorithm is selected. The loss parameter is the cost function used in the model training process. We choose the mean squared error cost function in this work. This cost function is suitable for regression problems. The metric parameter is the criteria used in the model training process to evaluate the model's performance. Accuracy criterion is chosen for this purpose. This measure is suitable for classification problems and shows what percentage of the test data is correctly classified.

At least, the accuracy and error diagram of the convolutional neural network is shown in Figure 15 and Figure 16, which are respectively related to the accuracy and error diagrams for the convolutional neural network model in the training process. The blue graph is related to the accuracy and error of the model for the training data and the orange graph is for the accuracy and error of the model on the validation data. As it is clear from the accuracy graph, in the initial stages of the model, it was able to achieve a relatively perfect fit of the data. So that both the accuracy of the model on the training data and the validation data have an accuracy of over 97.5% in each stage, and on the other hand, the error rate has reached below 1% in the same initial stages.

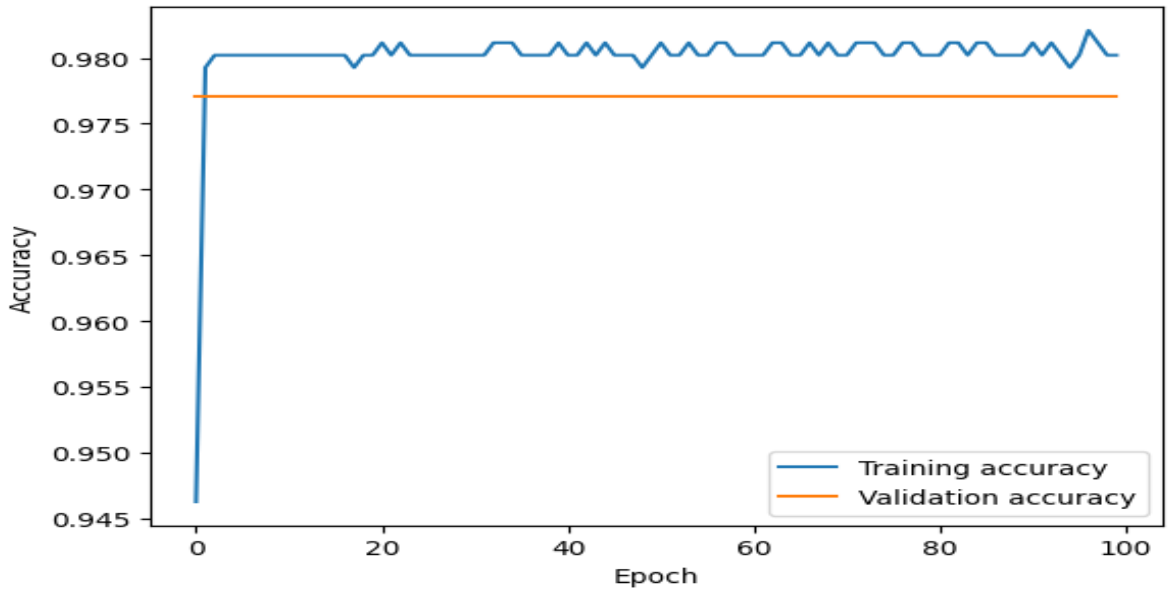

Figure 15: Accuracy of CNN

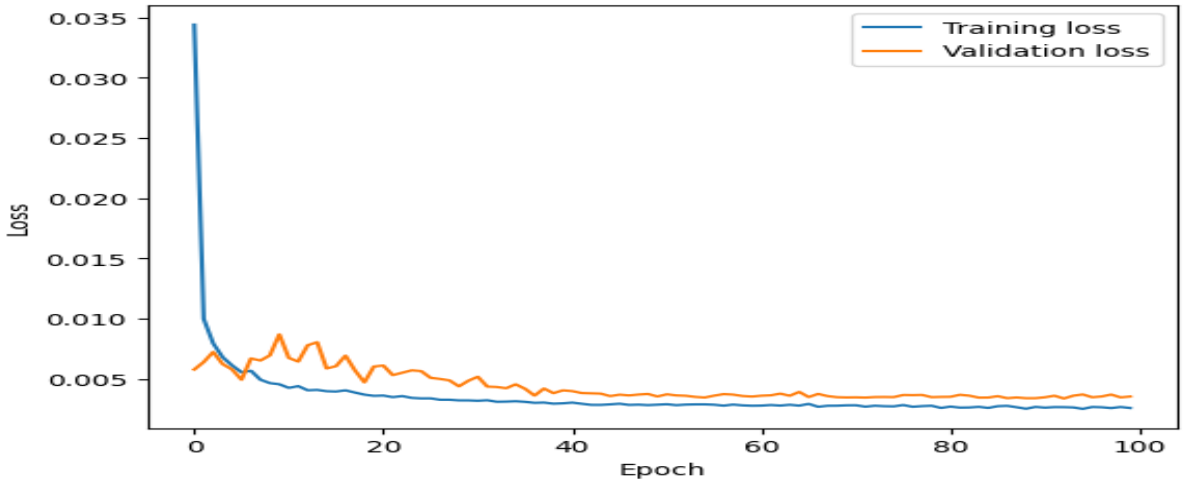

Figure 14: Error of CNN

3-7- combined of convolutional neural network and long-short term memory

The diagram of accuracy and error of the CNN-LSTM model in the training process is shown in Figure 17 and Figure 18, respectively. The blue graph is related to the accuracy and error of the model for the training data and the orange graph is for the accuracy and error of the model on the validation data. The CNN-LSTM model, unlike the CNN model, at the very beginning could not achieve a relatively perfect fit of the data in a stable manner, and after the 20th step of training, the accuracy of the model on the data increased and decreases every few steps, so that even In the last stages of training, the accuracy of the model on the training data decreases more. But overall, the model has been successful, because the lowest accuracy rate on the training and validation data was 97.85% and 97.70%, respectively. On the other hand, the error of the model has always been decreasing, so the amount of error has reached below 0.5% in the initial stages. Considering the accuracy of the model, this indicates that the model has not been over-fitted and finally a relatively complete model of the data has been fitted.

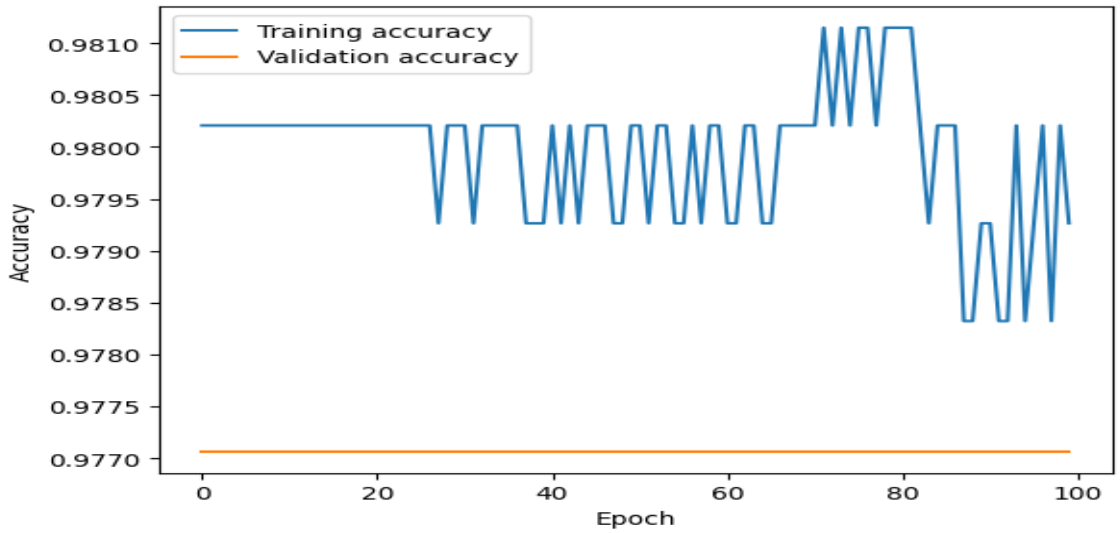

Figure 15: Accuracy in CNN-LSTM

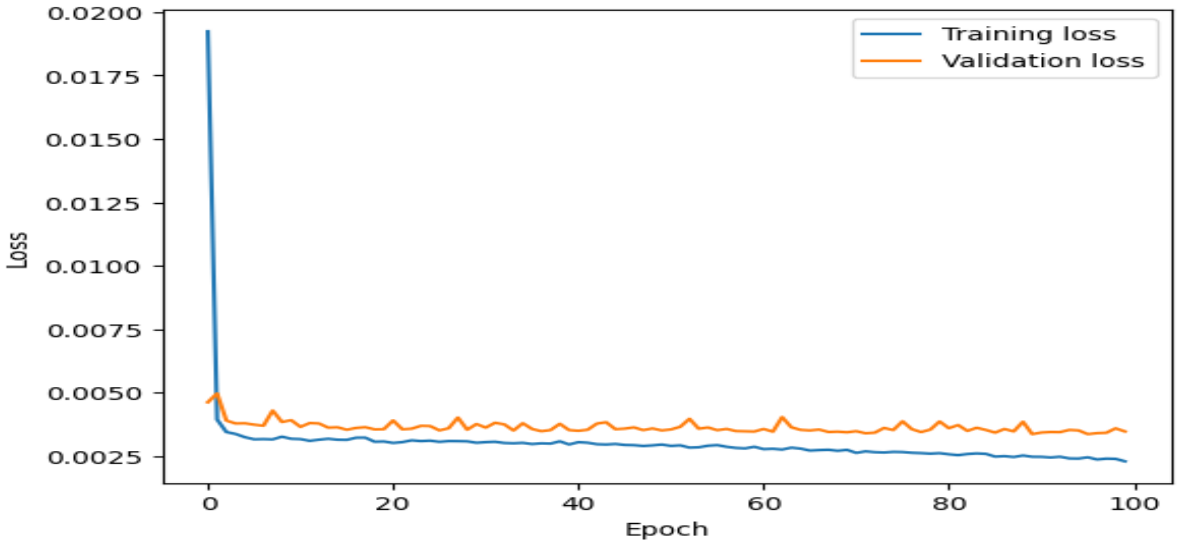

Figure 16: Error in CNN-LSTM

The accuracy of the CNN model on the test data is equal to 78.97%. The accuracy of CNN-LSTM model on the same test data is equal to 98.23%. Considering that both models were trained in the same conditions and the learning parameters of both were exactly equal, and considering the higher accuracy of the CNN-LSTM model on the test data, it can be concluded that the CNN-LSTM model is better than the CNN model has acted. But in general, both models have been able to obtain good accuracy on the test data.

3-8- Predicting the income and risk of each customer

As mentioned, the CNN-LSTM model was identified as the best model for predicting the risk and profit of customers requesting facilities, because the accuracy of the model in the test data reached 98.23%.

The amount of profit and risk of each cluster by using CNN-LSTM model are presented in the following table:

Table 3: amount of profit and risk of each cluster by using CNN-**LSTM**

cluster	Feature of customer in each cluster	Predict Income of each cluster	Predict Risk of each cluster
1	Female in Service Department	29786130	0.45
2	Female in Housing Department	2070888	9.71
3	Female in Commerce Department	0	0
4	Male in Service Department	3146007	55.6
5	Male in Housing Department	29786130	34.9
6	Male in Commerce Department	65879432-	38.33
7	Company in Service Department	758593429	0.45
8	Company in Housing Department	716548852-	24.56
9	Company in Commerce Department	11395659	8.23
10	Company in Industry Department	3347505699	0

As it can be seen in table number 3, the amount of the highest profit and zero risk belongs to the tenth cluster, which means that by paying the facility to the tenth group, the profit of the bank will be maximum and its risk will be minimized, so the companies in the industry sector are the least risky and The most profitable customers of banks are in terms of payment facilities. And then the payment of facilities to the companies in the service sector with the highest income and very low risk regarding the repayment of the facilities according to the above table is a priority. Anticipating the payment of facilities to companies in the housing sector and men in the commercial sector has costs for the bank.

4- Conclusion

In this article, the behavior of customers in the Iran banks has been investigated. In this article, Data Mining and Deep Learning are used for predicting the behavior of bank customers. Three types of feature selection algorithms are were compared with each other and the random forest algorithm was chosen as the best algorithm for clustering bank customers because it had selected the best features in both target variables. Customers are divided into two groups, Low-Risk and High-Risk customers. Banks don't want to pay facilities to bad credit customers, They pay facilities for good credit cluster. the best way to predict the amount of income and risk of each customer before payment is to use the CNN-LSTM algorithm. By identifying the amount of risk and income of each customer, the bank can make appropriate decisions to pay or not pay the facility and also predict the ability of the customer to repay of loan and bank can decides about appropriate amount of the facility to each client. The best amount payment of the facility is designed and the non-repayment of the facility is prevented.

Due to the fact that the research data was collected from one of the private banks in Iran, the researcher has reached these results, although the research method can be used, but using the data of other banks, different results may be obtained. In this research, the first limitation is the access to statistics and information because many branches of Bank have avoided giving personal information to the researcher because they consider it private. The lack of research studies related to the research topic is another limitation of this research .The subject of the research regarding the use of neural networks in deep learning has been used very little for bank data and these classification models have been used mostly for predicting pictures and words and contracting diseases. So far, the combined CNN-LSTM model has not been used for bank data to predict payment facilities at all

References

Adnan et al, R.M.,(2020). Least square support vector machine and multivariate adaptive regression splines for streamflow prediction in mountainous basin using hydro-meteorological data as inputs,J. Hydrol.156-159.

Al Machot,F. et al.(2019). A deep-learning model for subject-independent human emotion recognition using electrodermal activity sensors.219-231.

Ansari,A., Riasi.,A., (2016). Customer Clustering Using a Combination of Fuzzy C-Means and Genetic Algorithms, 34-51.

Armantier, O., Ghysels, E., Sarkar, A., Shrader, J., (2015). Discount window stigma during the 2007-2008 financial crisis. J. Financ. Econ. 118 (2), 317–335.

Bernanke, B.,(2009). The federal reserve's balance sheet: an update. In: A Speech at the Federal Reserve Board Conference on Key Developments in Monetary Policy, Washington, D.C.210-213.

Bhandari, P.,(2022). Missing Data | Types, Explanation, & Imputation. Revised on November 11, 2022.

Bock , H.,(2007). Clustering methods: A history of k-means algorithmsSelected Contributions in Data Analysis and Classification, . 161-172.

Bose, I., Chen, X.,(2015). Detecting the migration of mobile service customers using fuzzy clustering.

Chen, X.D. ,(2017). Analysis and research of common clustering algorithm in data miningDigital Technol. Appl., pp. 151-152.
De Andrés et al. J.,(2011). Bankruptcy forecasting: a hybrid approach using Fuzzy c-means clustering and Multivariate Adaptive Regression Splines (MARS). 45-51.
Dormann et al, C.F.,(2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance.201-203.
Du S., Li J.,(2019). Parallel processing of improved knn text classification algorithm based on Hadoop. In: 2019 7th International Conference on Information, Communication and Networks (ICICN), 167-170.
Estivill-Castro V., Houle M.E.(2001). Robust distance-based clustering with applications to spatial data mining,Algorithmica, 30 (2) , 216-242.
Gharachorloo, N., Nahr, J. G., & Nozari, H. (2021). SWOT analysis in the General Organization of Labor, Cooperation and Social Welfare of East Azerbaijan Province with a scientific and technological approach. International Journal of Innovation in Engineering, 1(4), 47-61.
Guo,G., Wang, H., Bell, D., Greer, Y.,(2003). KNN model-based approach in classificationOTM Confederated International Conferences on the Move to Meaningful Internet Systems, 986-996.
Hajjem, A., Bellavance, F., Larocque, D.,(2014). Mixed-effects random forest for clustered dataJ. Stat. Comput. Simul, 1313-1328.
Henderi, T., Wahyuningsih,T. ,Rahwanto,E.,(2021). Comparison of Min-Max normalization and Z-Score Normalization in the K-nearest neighbor(KNN) Algorithm to Test the Accuracy of Type so Breast Cance, International Journal of Informatics and Information SystemVol.4,NO.1,13-20.
Howell, E.,(2021). 4 Techniques To Deal With Missing Data in Datasets, Simple methods that can nullify the effects of missing values, Published in Towards Data Science, Sep 17.
Huang et al., Huang J., Wei Y., Yi J., Liu M.,(2018). An improved kNN based on class contribution and feature weighting. In: 2018 10th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), 313-316.
Huang, X.Y. ,(2018). An improved KNN algorithm and its application in real-time car-sharing predictionDalian University of Technology, Daian, China ,M.S. thesis.222-225.
Khosroyani, M., Heydarpoor, F., Yaghoob-nezhad, A., & Poorzamani,Z.(2022). An artificial neural network model for predicting the liquidity risk of Iranian private banks. Int. J. Nonlinear Anal. Appl. In Press, 1–11 ISSN: 2008-6822 (electronic); http://dx.doi.org/10.22075/ijnaa.2022.29118.4071. [In Persian].
Miller, A., (2002). Subset Selection in Regression, 2nd ed. Boca Raton, USA: Chapman and Hall, CRC Press.10.1201/9781420035933.
Mohammadi, N., Zangeneh, M.,(2016). Customer credit risk assessment using artificial neural networksIJ Information Technology and Computer Science, 8 (3) , 58-66.
Ngufor, C., Van Houten, H., Caffo, B.S., Shah, N.D., McCoy, R.G.,(2019). Mixed Effect Machine Learning: a framework for predicting longitudinal change in hemoglobin A1cJ. Biomed. Inform, 56-67
Nozari, H. (2024). Investigating Key Dimensions and Key Indicators of AIoT-Based Supply Chain in Sustainable Business Development. In Artificial Intelligence of Things for Achieving Sustainable Development Goals (pp. 293-310). Cham: Springer Nature Switzerland.
Nozari, H., Sadeghi, M. E., Eskandari, J., & Ghorbani, E. (2012). Using integrated fuzzy AHP and fuzzy TOPSIS methods to explore the impact of knowledge management tools in staff empowerment (Case study in knowledge-based companies located on science and technology parks in Iran). International journal of information, business and management, 4(2), 75-92.
Ole Hjelkrem,L., Eilif de Lange.P(2023). Explaining Deep Learning Models for Credit Scoring with SHAP: A Case Study Using Open Banking Data, J. Risk Financial Manag. 2023, 16(4), 221; https://doi.org/10.3390/jrfm16040221.
Ray,S., (2019). A quick review of machine learning algorithms. In: 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon), IEEE, 35-39.

Sabo, S., (2014). Analysis of the k-means algorithm in the case of data points occurring on the border of two or more clusters.

Speiser, J.L.,Wolf, B.J., Chung, D., Karvellas, C.J., Koch, D.G., Durkalski V.L.,(2019). BiMM forest: A random forest method for modeling clustered and longitudinal binary outcomesChemometrics Intell. Lab. Syst.

Tavakkoli-Moghaddam, R., Ghahremani-Nahr, J., Samadi Parviznejad, P., Nozari, H., & Najafi, E. (2022). Application of internet of things in the food supply chain: a literature review. Journal of applied research on industrial engineering, 9(4), 475-492.

Teles, G., Rodringues, J., Rabelo , R., & Kozlov, S.,(2020) Artificial neural network and Bayesian network models for credit risk prediction, J. Artif. Intel. Syst. no. 1, 118–132.

Wang, P.(2018). Research on Application of Big Data in Internet Financial Credit Investigation Based on Improved GA-BP Neural Network, Volume 2018, Article ID 7616537, https://doi.org/10.1155/2018/7616537.

Wang, Q., Wang, S., Wei, B., Chen, V., Zhang, Y.,(2021). Weighted K-NN classification method of bearings fault diagnosis with multi-dimensional sensitive featuresIEEE Access, 45428-45440, 10.1109/ACCESS.2021.3066489.

Wang,h.,Xu,P.,Zhao , J.,(2022). Improved KNN algorithms of spherical regions based on clustering and region division, Alexandria Engineering Journal,Volume 61, Issue 5, 3571-3585.

Yanenkova ,I .Nehoda,Y. Drobyazko,S. Zavhorodnii, A. Beresovska,L(2021). Modeling of Bank Credit Risk Management Using the Cost Risk Model, Risk Financial Manag. , 14(5), 211; https://doi.org/10.3390/jrfm14050211

Yang, Z., Xiaolong ,Su.,(2012). Customer Behavior Clustering Using SVM.

Zach,A.,(2021). Z-Score Normalization: Definition & Examples, https://www.statology.org/z-score-normalization/

Zhang, Q., Couloigner ,I.,(2005). A new and efficient K-medoid algorithm for spatial clustering. In: Lecture notes in computer science, vol. 3482, (III), http://dx.doi.org/10.1007/11424857_20.

Zhang,Z.,(2016),Introduction to machine learning: k-nearest neighbors,Ann. Transl. Med., 4 (11) .

Zhou, S., Bao, Y. Lu, D., Wang, K., Shan, J., Hou, Z., (2022). Real-time data-driven fault diagnosis of proton exchange membrane fuel cell system based on binary encoding convolutional neural network,Int J Hydrogen Energy, 47 (20) , 10976-10989.