International journal of industrial engineering and operational research (IJIEOR)
Volume 2, No. 1, 2020
www.ijieor.bgsarman.ir, www.ijieor.ir **ISSN: 2676-3311**

Assessing and validating bank customers using data mining algorithms for home loan

Reza Parvizi [1*] , Mohammad Amin Adibi[2]

[1] *Faculty of Industrial and Mechanical Engineering, Islamic Azad University, Qazvin Branch, Iran*
[2] *Faculty of Industrial and Mechanical Engineering, Islamic Azad University, Qazvin Branch, Iran*

Article info:

Received 2020/05/05
Revised 2020/05/18
Accept 2020/05/20

Keywords:
Data mining
Validating
Bank loan
Credit risk

Abstract

All over the world, the banking industry is one of the most important pillars of any country's economy and due to the provision of various financial and credit services, it plays a decisive role in the development and economic growth of the country and it can be described as the driving force that accelerates, balances, and organizes the economy. Given that banks' capital is less than the total value of their assets, even if a percentage of the loans are not receivable, banks face bankruptcy risk. In order to reduce the volume of claims, validation of loan applicants can be very effective. Data mining techniques and algorithms are one of the methods of validating bank customers. In this study, considering the current and past data of bank customers To receive loan facilities analyzed by data mining algorithms such as decision trees, K nearest neighbor algorithm (KNN), support vector machine (SVM) and Random Forest to identify good and bad customers. Finally, the ability of these algorithms to validate banking customers will be examined. According to the results, Random forest algorithm has a better ability to classify customers.

1. Introduction

Assessing and measuring the repayment capacity of credit customers individually and calculating the probability of non-repayment of the received credits is called "validation" [1]. Validation is a system by which banks and credit institutions use the applicant's current and past information which evaluate the possibility of non-repayment of the loan by the applicant [2]. This approach has been an objective tool for managing banks' credit risk and classifies credit customers as neutral and based on qualitative and quantitative statistics and information. In this regard, many methods have been used to classify facility applicants and it has always been the intention of researchers to increase the accuracy of these methods. Considering the remarkable capabilities of data mining techniques in classification and forecasting, today, the use of these techniques in issues related to risk assessment, validation, forecasting and other issues has received a lot of attention. Therefore, considering the importance of this issue and using data mining techniques, the risk of non-repayment of credits can be predicted and reduced. Using a variety of data mining techniques, the information obtained can be used to achieve goals like increased revenue, reduced costs, improved customer relationship and reduced risk. The purpose of this project is to design a mechanism and models to validate the bank's customers using

*Corresponding author email address: rezaparvizy@rocketmail.com

data mining techniques. Such algorithms has been used the assessment and validation of loan applicants at the level of individual loans for home purchase by loan applicants. This mechanism guides the credit risk management decisions and the bank's credit circle in order to identify and allocate loans to reputable customers. At the same time, it supports loan portfolio management decisions for optimal utilization of the portfolio's diversity and portfolio effects [3]. Risk forecasting leads to credit risk management decisions and the bank's credit circle, which guides reputable customers in identifying and allocating loans and based on the arrangement of new mechanisms, bank processes in the field of lending to customers will be reviewed. If possible, alternative processes are proposed to improve and accelerate the borrowing processes. Through such a mechanism, reputable customers are discredited and bank resources flow to applicants who are more eligible for a loan. The mechanism also regulates the guarantees and collateral required to lend to customers based on their credit and it prevents unnecessary rigidity and negligence towards applicants. In this way, customers with bank quality can borrow more easily and low quality customers can only get loans by providing strong guarantees.

After collection the data of the bank's customers from the relevant databases, identified variables that affected customer rankings then they used neural networks to classify the customer based on their characteristics. They provided a framework for predicting the credibility of new applicants and placing them in indebted and non-indebted classes [4].

Step by step regression method for selecting variables can only select specific variables and cannot select the same group of variables as a whole that cannot be interpreted. Using the lasso group can be a good solution to this problem, the results show that the lasso group method is better than other methods with predictive accuracy and can be useful and valuable in the development of credit validation model [5].

A way to predict whether a debtor is eligible for a loan, a combination method using the shopping cart algorithm and optimization of binary particles is presented. Several popular algorithms, such as neural network, logistic regression, and support vector machine, the proposed method showed outstanding performance [6].

Using Monte Carlo's experiments and combining several large sets of data, algorithms that estimate have been obtained. Through these techniques, financial companies provide banks with more loans when they anticipate that customers will be less risky in their payments. Of all the models, the generalized linear model had the most efficient and accurate calculations But the support vector machine was identified as the best way to predict credit risk [7].

Using the benefits of SVM and overcoming SVM based clustering methods, a completely new two-step classification method was introduced without supervision and there is no initial planning. The new second class SVM model is designed without monitoring, a classification algorithm suitable for balanced and unbalanced data. It has expanded the proposed approach to credit risk assessment [8].

The customer satisfaction and the amount customer expectations are met have maximum weight or priority in achieving efficiency among recognizing factors. The researcher investigated efficient units using a combination of fuzzy Data Envelopment Analysis (DEA) and fuzzy Analytic Hierarchy Process (AHP) [9].

All of researchers selected one or some algorithms to predict and measuring the credit risk by classification techniques. They introduced new algorithms to classify bank customers. They used techniques and algorithms as good as they could perform, however they could use other indicators to measure efficiency of algorithms. There are many different indicators which they can measure ability and efficiency of algorithms and techniques. In this research used some data mining algorithms to validate bank customers by using several indicators which can measure algorithm efficiency in bank costumer validating field in addition to Assessment and validating bank customers.

The methods and algorithms used by previous researchers in this field of research are as described in Table 1.

Table 1. Table of algorithms used over the past years

Algorithms/ source	Artificial neural networks	K nearest neighbor algorithm	Support vector machine	Lasso group	Logistic regression	Quadratic discriminant analysis	Decision tree (CART)	Prune decision tree	Linear regression mode	Linear mixed model	Generalized linear model	Data Envelopment Analysis	Quadratic surface support vector machine	Linear discriminant analysis (LDA)	Multivariate adaptive regression splines (MARS)	Genetic programming models	Random forest
Chen et al. [16]	1		1				1							1	1	1	
Kruppa et al. [17]		1			1												
Verbrake et al. [4]	1																
Hongme & Yaoxin [5]	1			1													
Malik [6]	1		1		1		1										
Pérez Martín et al. [7]	1		1			1	1	1	1	1	1						
Nazeri & Keshavarzi [9]												1					
Jian Luo et al. [8]													1				
Current research		1	1				1										1

In section 2, the research method, database, algorithms and their characteristics will be examined, in section 3, the research findings and the results obtained from the analysis and evaluation of the implementation of the algorithms will be examined. And in section 4, the results obtained from the analysis of data and algorithms will be examined.

2. Data base and algorithms

2.1. Data base

In this study, 13 characteristics of customers and applicants for home loans have been evaluated. The input parameters and data used in this study include table 2 items. The first 12 rows include the features and data of the loan applicants for evaluation. Row 13 is related to the final status or final class for the allocation or non-allocation of the loan to the applicant. 70% of database data is trained and 30% of them have been used as experimental data.

Features of table 1 data are common features used in credit rating models in most countries of the world such as New Zealand, Japan, Singapore, Germany, Portugal, Spain, Italy, as

evaluation indicators. These indicators include demographic, financial, and occupational and credit behavior indicators.

Table 2. Parametric table of loan applicant data

Row	Parameter name	Characteristic	Index type	Data type
1	Applicant age	Numerical	Demographic indicators	Quantitative
2	Gender	Male	Demographic indicators	Quality
		Female	Demographic indicators	Quality
3	Marital status	Married	Demographic indicators	Quality
		Single	Demographic indicators	Quality
4	The number of dependents	Number	Demographic indicators	Quantitative
5	Location feature	Urban	Demographic indicators	Quality
		Rural	Demographic indicators	
		Town	Demographic indicators	
6	Residential ownership status	Ownership	Demographic indicators	Quality
		Tenant	Demographic indicators	
7	Employment status	Employed	Occupation index	Quality
		Self employed	Occupation index	
8	Applicant's income	Number	Financial indicators	Quantitative
9	Applicant's side income	Number	Financial indicators	Quantitative
10	The amount of the loan requested	Number	Credit Behavior Indicators	Quantitative
11	Loan repayment period	Number	Credit Behavior Indicators	Quantitative
12	Application history by the applicant	Yes	Credit Behavior Indicators	Quality
		No	Credit Behavior Indicators	
13	Granting loans and facilities	Good	Category of applicants	Quality
		Bad	Category of applicants	

According to table 1, the first floor contains demographic indicators. These variables are not very important but are useful for obtaining gender and regional information. For instance, in terms of gender, older women who apply for a loan have a lower risk of repaying the loan than young male applicants. In general, the risk of non-repayment decreases with age. The risk is also lower for married applicants or applicants who have fewer dependents have a lower risk of repayment. It is also job indicator is very important to validate customers. Most people in developed countries have self-employed jobs and therefore earn less points than employees.

Revenue indicators include data on family income, job income and financial situation.

Behavioral indicators are variables that are examined for validating. If the customer has a shadow in the bank, the bank can make better decisions, because the bank can make a better decision by examining the current or past situation of the applicant. Loan repayment period and the amount of loan requested is another variable of this group that can be an indicator for customer evaluation.

In this research, decision tree, support vector machine, k nearest neighbors and random forest algorithms have been used to evaluate and validate customers. These algorithms can determine the final class of customers according to the large amount of database data and input variables.

2.2. Decision tree algorithm (DT)

The decision tree, which main purpose is to categorize data, is a model in data mining that, similar to flowchart, provides us like with a tree structure to make decisions

and determine the class and category of a particular data. The decision tree is made up of a number of nodes and branches in such a way that the leaves show the categories and the middle nodes are used to make decisions based on one or more specific attributes. This algorithm divides the data into specific sets, and each set of said sets is a subset of more or less homogeneous data that has predictable features [10]. If our problem is categorization and be $y_i \in [1,...,k]$. The impurity function for the Q node can be one of the 1 or 2 methods.

Formulas 1 and 2 parameters are as define:

p_i	possible of sample of data belongs to the class I
k	Number of classes in the training data

Gini's gross function (1).

$$Gini(Q) = 1 - \sum_{i=1}^{k} p_i^2 \qquad p_i = (p_1, p_2, p_3, ..., p_n) \tag{1}$$

Entropy gross function (2).

$$Entropy(Q) = -\sum_{i=1}^{k} p_i \log p_i \qquad p_i = (p_1, p_2, p_3, ..., p_n) \tag{2}$$

The diagram of the decision tree algorithm is Figure 1.

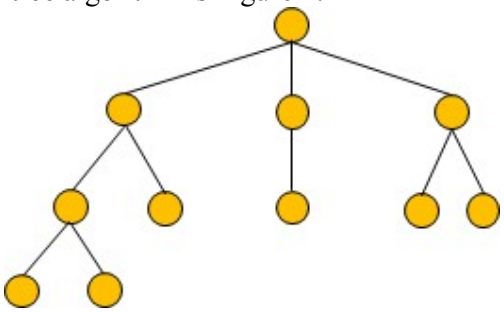

Figure 1. Decision tree algorithm diagram

Figure 1 showed some circles and arrows. Circles are nodes which actually they are attributes and arrows connect the attributes to predict the final class.

2.3. The K nearest neighbor algorithm (KNN)

The KNN algorithm is an optimization problem for finding the closest points in metric spaces. This algorithm is a non-parametric statistical method used for statistical classification and regression. This algorithm categorizes a test sample based on a close neighbor. Educational examples are presented as vectors in the multidimensional property space. The space is divided into areas with educational examples of partitioning. A point in space belongs to a class that has the most instructional points belonging to that class inside the closest instructional example to k. In the KNN algorithm, the closest neighbor to a sample is categorized by a majority vote of its neighbors. In this case, the class closest to K is measured by its closest neighbors using the distance function. The case is then assigned to the nearest neighbor's class [11].
The distance function is calculated as one of the 3, 4 or 5 methods.

Formulas 3, 4 and 5 parameters are as define:

x_i	Coordinates of x_i

y_i Coordinates of y_i
p positive integer

Euclidean distance function (3).

$$y = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \qquad x_i = (x_1, x_2, x_3, ..., x_n),\ y_i = (y_1, y_2, y_3, ..., y_n) \tag{3}$$

Manhattan distance function (4).

$$y = \sqrt{\sum_{i=1}^{k} |x_i - y_i|} \qquad x_i = (x_1, x_2, x_3, ..., x_n),\ y_i = (y_1, y_2, y_3, ..., y_n) \tag{4}$$

Minkowski distance function (5).

$$y = \left(\sum_{i=1}^{k} (|x_i - y_i|)^p\right)^{\frac{1}{p}} \qquad x_i = (x_1, x_2, x_3, ..., x_n),\ y_i = (y_1, y_2, y_3, ..., y_n),\ p = 1, 2, 3, ... \tag{5}$$

The diagram of the KNN algorithm is Figure 2.

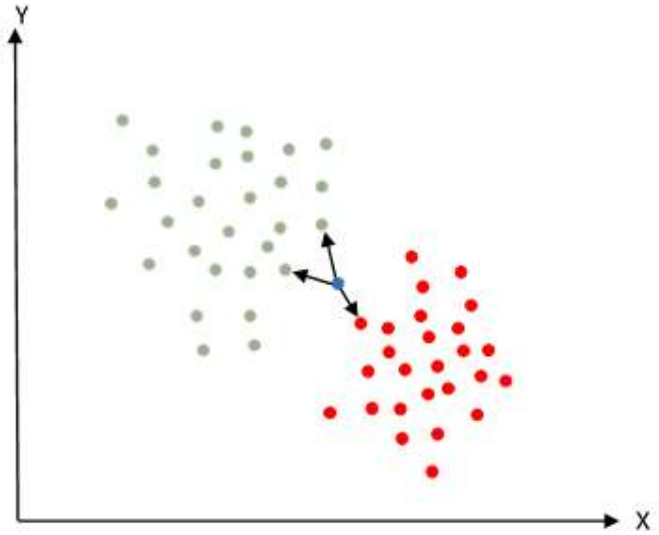

Figure 2. KNN algorithm diagram

Figure 2 is showing some green and red circles and a blue circle which has some arrows connected to green and red circles. Green and red Circles are two different classes and blue circle is a data which should classify to one of blue or red class.

2.4. Support vector machine algorithm (SVM)

The support vector machine is one of the supervised learning methods used for classification and regression. The backup vector machine, or SVM, is actually a Two-sided classifier. The SVM method tries to create a cloud page for two floors with a maximum distance between each floor and the super page. The point data closest to the cloud page is used to measure this distance. Hence, these point data are called backup vectors [12].
The SVM algorithm is classified as a pattern recognition algorithm. The SVM algorithm can be used wherever there is a need to identify patterns or categories of objects in specific

classes [13]. The SVM solution and modeling method is in the form of phrase 6, which is a non-linear programming optimization problem.

Formulas 6 parameters are as define:

x_i	Backup vectors
w	Vectors with a number of members
b	Numerical constant value
$\frac{1}{2}\|w\|^2$	The distance between the separator superhero

Mathematical modeling of support vector machine (6).

$$y = Min\frac{1}{2}\|w\|^2$$
$$s.t.$$
$$y_i(w.x_i - b) \geq 1 \qquad \forall 1 \leq i \leq n \qquad (6)$$

The diagram of the SVM algorithm is Figure 3.

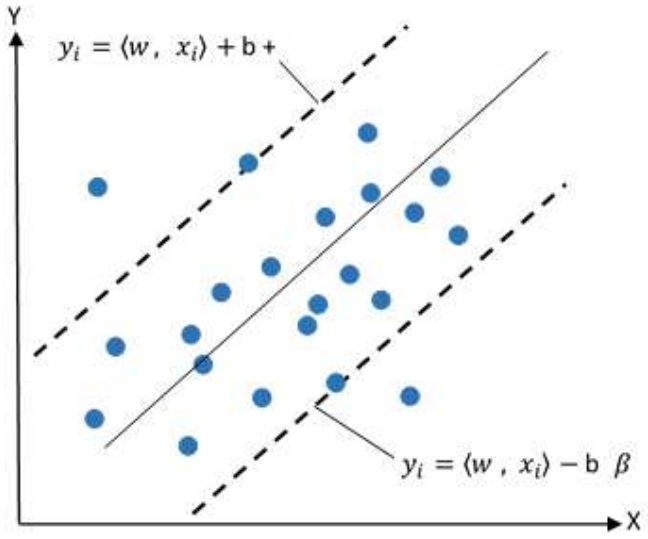

Figure 3. SVM algorithm diagram

Figure 3 showed three lines. Two of them are stretch lines and the middle one is continuous line. Continuous line separate data to two different classes.

2.5. Random forest algorithm

Random forests are a combined learning method for categorization, regression. Based on the structure of a large number of decision trees, they work on the time of training and output of classes (classification) or for predictions of each tree separately. Random forest as its name implies, the decision tree algorithms randomly generates. The "forest" is actually a group of decision trees. Random forest algorithms use decision trees for their simple and weak algorithms. A decision tree algorithm can easily perform classification operations on data, while several random decision trees are used in a random forest algorithm. In fact, a set of decision trees together produce a forest, and this forest can make better decisions than a tree. There are differences between the decision tree and the random

forest. If an input data set with different attributes is given as an input to the algorithm, some sets of rules are formulated in such a way that they are used to predict. In comparison, the tree algorithm randomly selects observations, decides on the properties of multiple trees, and then calculates the average of the results [14].
Formulas 7 parameters are as define:
The function of the random forest is in the form of formula 7.

B	Number of trees
f_b	Output the ensemble of trees
x'	make a prediction at a new point

$$y = \frac{1}{B}\sum_{b=1}^{B} f_b(x') \qquad B = (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) \ , \ f_b = \{(x_i, y_i)\} \tag{7}$$

The random forest diagram is Figure 4.

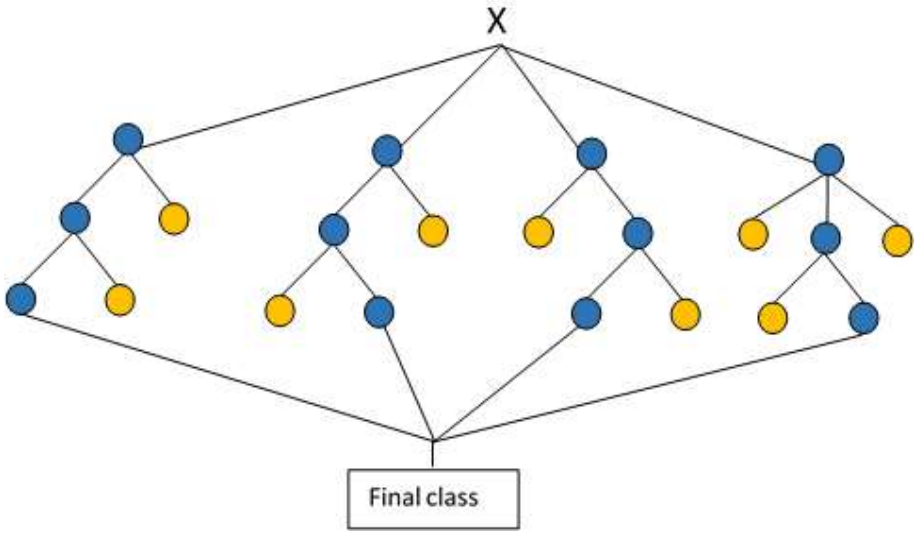

Figure 4. Random forest algorithm diagram

Figure 4 showed four decision trees. Random forest consist of several decision trees. Blue circles are optimization nodes and classes which has chosen by each tree. Combining all of decision trees make decision for final class.

2.6. Algorithm performance indicators

Functional indicators in the implementation of algorithms include indicators by which the ability to measure and predict algorithms can be measured. The most important of these indicators include accuracy, MSE, RMSE, AUC, recall, precision, and K-Fold. These indicators are a good assessment for the validation of bank customers. These criteria can be calculated both for the training data set at the learning stage and for the test record set at the assessment stage. Evaluation in classification algorithms is done using the classification matrix concept.

Table 3. Classification Matrix

		Ptrdicted	
		Negative	Positive
Actual	True	True nagative (TN)	False Positive (FP)
	False	False nagative (FN)	True Positive (TP)

Each of the matrix elements is as follows:

TN: It represents the number of records which real category is negative, and the classification algorithm has correctly identified their category.
TP: It represents the number of records which actual category is positive and the category algorithm has correctly identified their category.
FP: It represents the number of records which real category is negative and the category categorization algorithm has mistakenly identified them as positive.
FN: It represents the number of records which actual category is positive and which category categorization algorithm has mistakenly identified negative.

The accuracy Index indicates the number of correct predictions made by the category, divided by the total number of predictions made by the same category.
The accuracy index formula is function (8).

$$accurcy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

The recall criterion is the ratio of the number of correctly classified data in a particular class to the total number of data that must be classified in that particular class.
The recall index formula is function (9).

$$recall = \frac{TP}{TP + FN} \tag{9}$$

The precision criterion evaluates the ratio of the number of correct predictions made for samples of a particular class to the total number of predictions for samples of that particular class.
The precision index formula is function (10).

$$precision = \frac{TP}{TP + FP} \tag{10}$$

The MSE index distinguishes between estimated values and what is estimated. The closer it is to zero, the lower the error rate.
The precision index formula is function (11).
Formulas 11 parameters are as define:

x_i	Equivalent to system output
y_i	Equivalent to definitive answer
n	Number of records

$$MSE = \frac{1}{n}\sum_{i=1}^{k}\left|x_i - y_i\right|^2 \qquad x_i = (x_1, x_2, x_3, ..., x_n),\ y_i = (y_1, y_2, y_3, ..., y_n) \tag{11}$$

The comprehensive k-Fold Cross validation method of the whole data set is divided into k equal parts. The k-1 part is used as a set of training data. Based on that, the model is made and the evaluation is performed with the remaining part. This process will be repeated as many times as k. The process will be repeated as many times as k, so that each k part is used only once for the evaluation, and each time an accuracy is calculated for the model being made. In this method, the final accuracy of the classification will be equal to the calculated k accuracy. The most common value in scientific texts for k is 10.

The AUC represents the area under the chart (ROC). The larger the number of this category in a category, the more optimal the final performance of the category. The ROC chart is a way to check the performance of categories.
The AUC index formula is a function of (12).

$$
\begin{aligned}
&1 \rightarrow TPR = \frac{TP}{TP + FN} && TPR \rightarrow y\,(axis) \\
&2 \rightarrow FPR = \frac{FP}{FP + TN} && FPR \rightarrow x\,(axis) \\
&1,2 \rightarrow AUC = \int_0^1 TPR\,(FPR^{-1}(x))dx
\end{aligned}
\tag{12}
$$

3. Results and Discussion

Customer validation in this research was first performed with the KNN algorithm, then DT algorithms, SVM, and finally Random forest. Analysis has done by python software.

3.1. KNN Algorithm results

The results obtained from the validation of bank customers by the KNN algorithm are as follows. The input parameters of the KNN algorithm are shown in Table 4.

Table 4. Input parameters of KNN algorithm

Parameter	Input
Input function	Minkowski distance function
K-number	5
Input attribute	Table a

The results of the classification matrix are from the implementation of the DT algorithm in the form of figure 5.

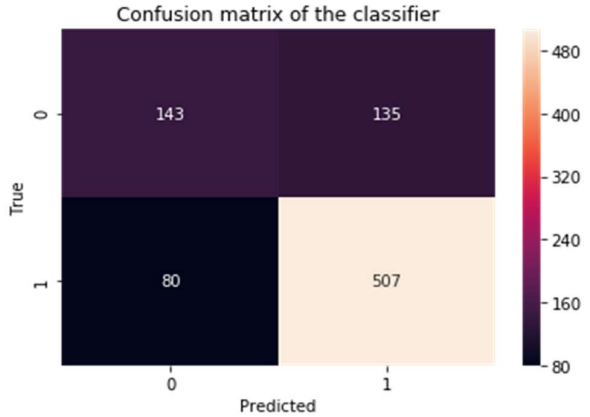

Figure 5. Classification matrix results of KNN algorithm

Figure 5 showed performance of KNN algorithm in confusion matrix. According matrix results, 507 members of data test predicted as true positive, 135 members predicted false positive, 143 members predicted true negative and 80 members predicted false negative.

3.2. DT Algorithm results

The results obtained from the validation of bank customers by the DT algorithm are as follows. The input parameters of the DT algorithm are as described in Table 5.

Table 5. Input parameters of DT algorithm

Parameter	Input
Input function	Gini's gross function
Min samples leaf	1
Input attribute	Table a

The results of the classification matrix are from the implementation of the DT algorithm in the form of figure 6.

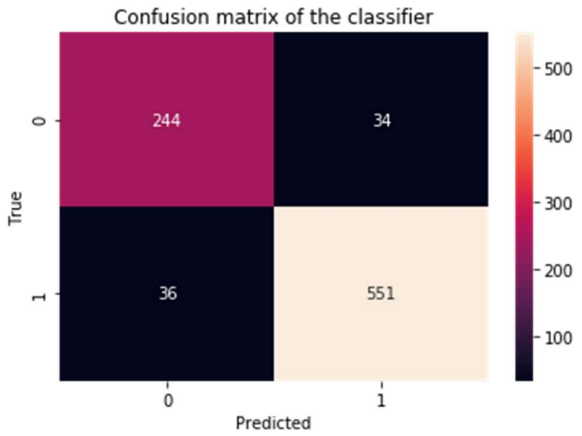

Figure 6. Classification matrix results of DT algorithm

Figure 6 showed performance of DT algorithm in confusion matrix. According matrix results, 551 members of data test predicted as true positive, 34 members predicted false positive, 244 members predicted true negative and 36 members predicted false negative.

3.3. SVM Algorithm results

The results obtained from the validation of bank customers by the support vector machine algorithm are as follows. The input parameters of the support vector machine algorithm are as described in Table 6.

Table 6. Input parameters of the SVM algorithm

Parameter	Input
Kernel	Linear
C	-
Input attribute	Table a

The results of the classification matrix are from the implementation of the support vector machine algorithm in the form of figure 7.

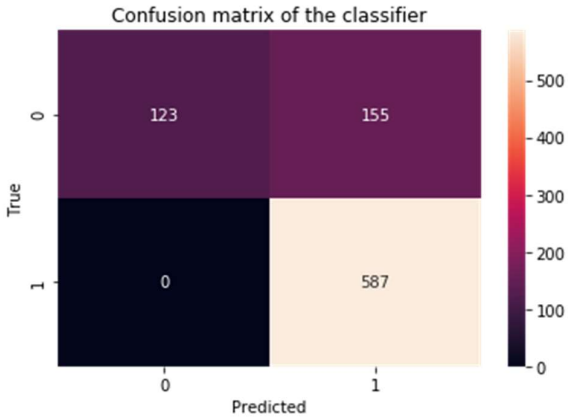

Figure 7. Classification matrix results of SVM algorithm

Figure 7 is showed performance of SVM algorithm in confusion matrix. According matrix results, 587 members of data test predicted as true positive, 155 members predicted false positive, 123 members predicted true negative and 0 members predicted false negative.

3.4. Random forest Algorithm results

The results obtained from the validation of bank customers by random forest algorithm are as follows. The input parameters of the random forest algorithm are as described in Table 7.

Table 7. Input parameters of the random forest algorithm

parameter	input
Input function	Gini's gross function
estimators	10
input attribute	Table A

The results of the classification matrix are from the implementation of the random forest algorithm in the form of figure 8.

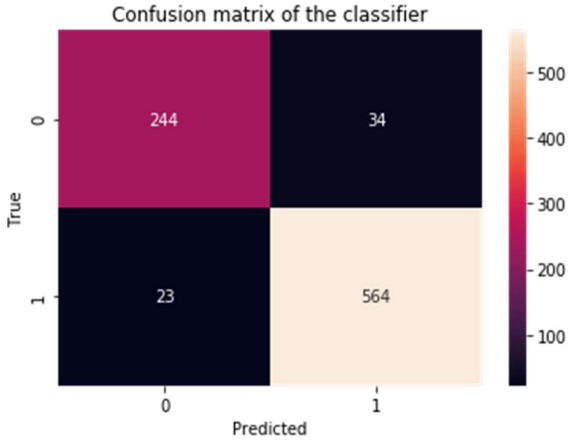

Figure 8. Classification matrix results of random forest algorithm

Figure 7 showed performance of SVM algorithm in confusion matrix. According matrix results, 564 members of data test predicted as true positive, 34 members predicted false positive, 244 members predicted true negative and 23 members predicted false negative.

The results obtained from the implementation of KNN, DT, SVM and random forest algorithms are as follows in table 8.

Table 8. Random forest algorithm results

indicators	Results			
	KNN	**DT**	**SVM**	**Random forest**
Accuracy	0.75144	0.91907	0.82080	0.93410
Recall	0.86371	0.93867	1.0	0.96081
Precision	0.78971	0.94188	0.79110	0.94314
MSE	0.49855	0.28447	0.42330	0.25670
AUC	0.863	0.947	0.972	0.972
Mean validation accuracy	0.79464	0.94347	0.76536	0.94692
1 of k fold 10 - (Accuracy-score)	0.75778	0.93425	0.75778	0.93771
2 of k fold 10 - (Accuracy-score)	0.78546	0.91695	0.75086	0.94117
3 of k fold 10 - (Accuracy-score)	0.81660	0.95155	0.74740	0.96193
4 of k fold 10 - (Accuracy-score)	0.79584	0.95501	0.77854	0.95501
5 of k fold 10 - (Accuracy-score)	0.83737	0.93079	0.73010	0.94463
6 of k fold 10 - (Accuracy-score)	0.80208	0.95138	0.75	0.95138
7 of k fold 10 - (Accuracy-score)	0.80555	0.95138	0.75694	0.94444
8 of k fold 10 - (Accuracy-score)	0.77430	0.93755	0.81184	0.94791
9 of k fold 10 - (Accuracy-score)	0.79790	0.94425	0.76041	0.95470
10 of k fold 10 - (Accuracy-score)	0.77351	0.96167	0.80968	0.93031

4. Conclusion

Given that credit risk exposures remain a major source of problems for the world's leading banks, banks and their inspectors need to be able to learn useful lessons from past business. Just as banks need to invest enough to compensate for the risks they face, they need to be aware of the need to identify, measure, care for, and control credit risk. Today, the risk management committee has been formed at the level of macro-management of banks and provides solutions in this regard. Credit risk management involves identifying, measuring, monitoring, and controlling risk, in other words, after identifying risk, it is measured using a variety of techniques and models, including multivariate regression, audit analysis, logit and project analysis, and neural network modeling. The more accurate the measurement, the better the monitoring and control of the risk in the next steps. Therefore, in order to increase the satisfaction of the applicants, the bank should make decisions and organize strategies and tactics dynamically using the exploration of the extracted data. The study used data mining algorithms such as the decision tree, the KNN, the support vector machine, and the random forest to validate bank customers. According to the results of the implementation of algorithms to explore the information data of home loan applicants, all four algorithms were able to have good results in customer accreditation. However, since it is very difficult to validate customers in large amounts of big data and the accuracy and power of measuring models has always been important for bank managers to make decisions, in this study we tried to validate customer validation by testing several data mining algorithms. The bank will be tested. According to the results, the random forest algorithm performed better for validation. The results in terms of measurement indicators showed that this algorithm had less error in validating according to the criteria. Therefore, this algorithm has a good ability to validate bank customers.

Other ways for solving this problem is mixed integer linear programming (MILP), in recent years, several mixed integer linear programming (MILP) models have been proposed for finding the most efficient decision-making unit in data envelopment analysis [15].

5. References

[1] Duffie, D., & Singleton, K. J. (2012). Credit risk: pricing, measurement, and management. Princeton university press.
[2] Herring, R. J. (2002). The Basel 2 approach to bank operational risk: Regulation on the wrong track. Journal of Risk Finance, 4(1), 42-46.
[3] Robert Jepsen, (2001), Modeling Data for Marketing, Risk, and Customer Relationship Management, 3rd edition, Wiley, New York
[4] Verbraken, T., Bravo, C., Weber, R., & Baesens, B. (2014). Development and application of consumer credit scoring models using profit-based classification measures. European Journal of Operational Research, 238(2), 505-513.
[5] Hongmei Chena, Yaoxin Xianga. (2017).The Study of Credit Scoring Model Based on Group Lasso, Procedia Computer Science 122 (2017) 677– 684
[6] Malik, R. F. (2018). Credit Scoring Using CART Algorithm and Binary Particle Swarm Optimization. International Journal of Electrical & Computer Engineering (2088-8708), 8.
[7] Pérez-Martín, A., Pérez-Torregrosa, A., & Vaca, M. (2018). Big Data techniques to measure credit banking risk in home equity loans. Journal of Business Research, 89, 448-454.
[8] Luo, J., Yan, X., & Tian, Y. (2020). Unsupervised quadratic surface support vector machine with application to credit risk assessment. European Journal of Operational Research, 280(3), 1008-1017.
[9] Nazeri, A., & Keshavarzi, M. (2019). Assessing the Performance of Branches of Refah Bank in Tehran Province by Combining Analytic Hierarchy Process (AHP) and Data Envelopment Analysis (DEA) Algorithms in Fuzzy Conditions. International Journal of Industrial Engineering and Operational Research, 1(1), 11-27. Retrieved from http://bgsiran.ir/journal/ojs-3.1.1-4/index.php/IJIEOR/article/view/4
[10] Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.
[11] Thirumuruganathan, S. (2010). A detailed introduction to K-nearest neighbor (KNN) algorithm. Retrieved March, 20, 2012.
[12] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.
[13] Dzulkifli, S.A., Salleh, M.N., & Talpur, K.H. (2019). Improved Weighted Learning Support Vector Machines (SVM) for High Accuracy. CIIS 2019.
[14] Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. In Ensemble machine learning (pp. 157-175). Springer, Boston, MA.
[15] Akhlaghi, R., & Rostamy-Malkhalifeh, M. (2019). A linear programming DEA model for selecting a single efficient unit. International Journal of Industrial Engineering and Operational Research, 1(1), 60-66. Retrieved from http://bgsiran.ir/journal/ojs-3.1.1-4/index.php/IJIEOR/article/view/12
[16] Chen, W., Xiang, G., Liu, Y., & Wang, K. (2012). Credit risk Evaluation by hybrid data mining technique. Systems Engineering Procedia, 3, 194-200.
[17] Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. Expert Systems with Applications, 40(13), 5125-5131.