Hindawi
Scientific Programming
Volume 2019, Article ID 8706505, 8 pages
https://doi.org/10.1155/2019/8706505

Research Article

Comparison of Data Mining Classification Algorithms Determining the Default Risk

Begüm Çığşar and Deniz Ünal

Cukurova University, Faculty of Arts and Sciences, Department of Statistics, Adana, Turkey

Correspondence should be addressed to Deniz Ünal; dunal@cu.edu.tr

Received 5 November 2018; Accepted 27 December 2018; Published 3 February 2019

Academic Editor: Antonio J. Peña

Copyright © 2019 Begüm Çığşar and Deniz Ünal. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Big data and its analysis have become a widespread practice in recent times, applicable to multiple industries. Data mining is a technique that is based on statistical applications. This method extracts previously undetermined data items from large quantities of data. The banking and insurance industries use data mining analysis to detect fraud, offer the appropriate credit or insurance solutions to customers, and better understand customer demands. This study aims to identify data mining classification algorithms and use them to predict default risks, avoid possible payment difficulties, and reduce potential problems in extending credit. The data for this study, which contains demographic and socioeconomic characteristics of individuals, were obtained from the Turkish Statistical Institute 2015 survey. Six classification algorithms—Naive Bayes, Bayesian networks, J48, random forest, multilayer perceptron, and logistic regression—were applied to the dataset using WEKA 3.9 data mining software. These algorithms were compared considering the root mean error squares, receiver operating characteristic area, accuracy, precision, F-measure, and recall statistical criteria. The best algorithm—logistic regression—was obtained and applied to the real dataset to determine the attributes causing the default risk by using odds ratios. The socioeconomic and demographic characteristics of the individuals were examined, and based on the odds ratio values, the results of which individuals and characteristics were more likely to default, were reached. These results are not only beneficial to the literature but also have a significant influence in the financial industry in terms of the ability to predict customers' default risk.

1. Introduction

The rapid and inevitable development of technology is causing a substantial global increase in the volume of data. Such data mean better information, and information is wealth. This is because information makes it possible for mankind to have a safer and better future, which is the primary goal of scientists and researchers. Due to this incredible amount of information that can be obtained from Big Data, humanity is able to make considerable progress in diverse fields ranging from health and safety to education and economy.

Obtaining information from big data utilizing the appropriate methods is similar to extracting the maximum possible ore from a newly discovered mine. The necessity of coming to scientifically accurate conclusions highlights the need for big data analysis. Big data analysis can reduce information loss and save time, giving rise to the term data mining (DM) [1, 2]. DM is a data analysis technique based on statistical application; it aims to extract information that could previously not be determined, from massive quantities of data [3].

Big data is not only a subject of interest for researchers but has also become an essential tool in business. Processing big data effectively is crucial for companies aiming for a leading role in their field. The need for big data analysis has especially increased in the banking and insurance industries. Even the small amount of information that has brought companies to the forefront of the competitive market, thanks to DM analysis, has increased the importance of DM.

The analysis and modeling of big data are not new subjects for actuaries, bankers, and insurers; DM helps them overcome many difficulties in their aim to manage money more effectively, control the system, reduce or transfer potential risks, understand client requirements, improve

funds management, increase market share, and reduce or transfer potential risks [4]. Specifically, DM can be used in the banking and insurance industries to determine default risks and risk groups, specify the correct insurance options for individual customers, increase customer satisfaction, and identify credit card fraud.

There are many DM methods to detect problems faced by bankers and insurers, for example, clustering, classification, and association. Classification is a widely used DM method that is applied in various fields [5]. Hence, the classification algorithms are widely used, and successful results obtained from the algorithms are also used for determining credit risks. Which classification algorithm to choose is a very important decision. There is no specific classification algorithm to solve the current problem. In other words, the best algorithm does not solve every problem in the best way. There are classification algorithms that give different results for different datasets or different problems. A classification algorithm that is considered to be the best solution to solve a problem may not work in another problem or dataset. For this reason, different classification algorithms for the given dataset must be compared before problem solving. The algorithm that best solves the problem is the algorithm obtained by comparison with the specific statistical criterions. Thus, the algorithm to be used in problem solving is determined. In this study, for determining best algorithms for current dataset, all data mining classification algorithms were compared with respect to the suitability of data and accuracy rates (accuracy threshold was taken as 80%). With this comparison, all algorithms were reduced to six classification algorithms (Naive Bayes, Bayes network, J48, random forest, multilayer perceptron, and logistic regression) that have almost the same accuracy threshold rate. Reduced algorithms—Naive Bayes, Bayesian networks, J48, random forest, multilayer perceptron, and logistic regression—are frequently seen algorithms in the literature. The six classification algorithms have almost the same accuracy rates and data availability. So, in order to determine the algorithm that will operate at the maximum level with the data, the comparison under various criteria was repeated using WEKA (Waikato Environment for Knowledge Analysis) 3.9 data-mining software.

The algorithms that have similar accuracy rates were compared again with different statistical criteria such as ROC (receiver operating characteristic), precision, recall, F-measure, and the root mean squared error (RMSE) to achieve the best results. As a result, the most appropriate algorithm for this dataset is found as the logistic regression algorithm.

The aim of this study is to use DM classification algorithms to investigate the effects of certain demographic and socioeconomic characteristics on the probability of individuals' default risk, as well as to predict their future payment challenges by determining individual attributes using a logistic regression classification algorithm.

The data for this study, which contain the demographic and socioeconomic characteristics of individuals, were obtained from the 2015 TUIK (Turkish Statistical Institution) survey. The raw data contained 59,663 observations, of which only 20,275 observations and 12 attributes belonging to heads of households were used. The data include the demographic features of households from various regions, as well as their total income and debts paid on a regular basis over the last 12 months.

When choosing an algorithm, using the algorithm which is known as giving good results without comparing its performance to different algorithms or comparing different algorithms by adhering to a single statistical criterion may give misleading results. Using an algorithm that is considered to only give good results without considering its suitability in data may lead to inaccurate results. Likewise, when determining the best algorithm, comparing under a single criterion may cause a wrong selection. Therefore, in this study, Naive Bayes, Bayes network, J48, logistic regression, multilayer perceptron, random forest and classification algorithms were determined with pre-elimination and then, determined algorithms were compared again with accuracy, F-measure, Roc area, recall, precision, and RMSE criterions. As a result of the analysis, the logistic regression classification algorithm was determined as the best algorithm. In conclusion, the logistic regression algorithm was used in the analysis of default risk.

Moreover, by applying the best algorithm (logistic regression) to the dataset, we determined which characteristics increase the default risk most.

2. Materials and Methods

The concept of extending credit goes back 5000 years and is still a primary research topic in the finance sector [6]. The role of banks and banking activities are expanding daily, which increases the necessity to manage loan issues appropriately. Currently, credit score models and credit ratings are used to determine an individual's default risk. Credit scores are mathematical models that determine the likelihood of a default risk by observing the characteristics of the customers applying for the loan [7].

In a loan society or company, crediting refers to the risk of balance at a certain time [8]. Credit institutions face many risks such as delays in payment or client defaults, the volatility of interest rates, and the depreciation of investments and securities. Credit risk management provides the opportunity to determine and measure these potential risks [9, 10].

Financial institutions design models using certain customer characteristics (age, gender, area of residence, income, marital status, previous credit payments, etc.) to predict and identify possible credit risks [11, 12]. Fisher proposed discriminant and classification analysis in 1936 as the basis of credit scoring models. Lately, decision trees, logistic regression, K-nearest neighbor, neural networks, and support vector machine algorithms are frequently used for credit scoring [13].

Financial institutions and analysts are always aiming to increase credit volume while reducing default risks. Therefore, credit scoring analyses are crucial to aid faster decision making, reduce the costs of loan analysis, monitor existing accounts more closely, predict default risks, and ensure that institutions can detect possible risks while

developing their competitiveness [14]. Therefore, DM techniques using big data should be applied for credit scoring [11].

In the last decade, there has been a significant increase in loan applicants and credit card users, which in turn has increased the risks for credit institutions. It is therefore necessary for banks and financial institutions to determine the probability of default risk by using the demographic and socioeconomic characteristics of customers. This allows financial institutions to take precautions against client default and identify risk groups. Identifying risk groups can also prevent potential customer losses and aid banks in avoiding potential risks. For this reason, DM using WEKA software is implemented to identify risk groups and ensure that financial institutions extend credit to clients not at risk of default.

The first part of our study compares the classification algorithms to select the most suitable algorithm according to selected criteria. In the second part, the best algorithm—the logistic regression—is used to research the attributes that may cause default risk. For this analysis, odds ratios were used as a criterion.

2.1. WEKA. The WEKA data-mining implementation software was developed by the University of New Zealand. It is an open source software program written in Java under General Public License. It contains several supervised and unsupervised methods such as classification, clustering, association, and data visualization. For this study, the WEKA 3.9 implementation and its experimenter user interface were used for the classification of the algorithms as well as to specify risk attributes using the logistic regression algorithm.

2.2. Dataset. The data for this study were obtained from the TUIK survey for 2015. Only heads of households over the age of 15 were selected from the 59,663 units of survey data. The data used to support the findings of this study are restricted by TUIK in order to protect privacy. The author(s) can supply the data upon request to researchers who meet the criteria for access to confidential data. The data contains the demographic and socioeconomic characteristics of individuals. A WEKA preprocessing application was used to obtain 20,275 units of data containing 12 variables, one of which is a class variable. Incidents of payment and nonpayment of past credit card debts are treated as class variables. The dataset structure is shown in Table 1.

Table 1: Dataset structure.

Attribute name	Description
Age	Age
Gender	Gender (1 = male, 2 = female)
Marital status	Marital status (1 = married, 2 = other)
Education	Education level (1 = illiterate, 2 = primary school, 3 = secondary school, 4 = high school, 5 = higher degree)
Work	Working status (1 = working, 2 = looking for a job, 3 = retired, 4 = other (nonactive))
Health	Health (1 = good, 2 = medium, 3 = poor)
Region	Region (1 = mediterranean, 2 = aegean, 3 = marmara, 4 = black sea, 5 = central anatolia, 6 = eastern anatolia, 7 = southeastern anatolia)
Housing	Housing status (1 = paying rent, 2 = not paying rent)
Revenue	Individual revenue (1 = low income, 2 = medium income, 3 = higher income)
Home loan	Nonpayment of house rent, interest-bearing debt repayment, or home loan payment within the last 12 months (1 = no, 2 = yes)
Bills	Nonpayment of electricity, water, and gas bills within the last 12 months (1 = no, 2 = yes)
Class	Nonpayment of credit card installments and other debt payments within the last 12 months (1 = no, 2 = yes)

2.3. Classification of Algorithms. Each object in the dataset is classified according to its similarities. Classification is the best-known and most used method of DM. The aim of the classification method is to predict accurately the target class of objects of which the class label is unknown [15]. In the WEKA implementation, classification algorithms are provided in nine groups. For the purposes of this study, the following algorithms were chosen: under the Bayesian file, Bayes networks (BayesNet) and Naive Bayes algorithms; under the Functions file, logistic regression (Logistic) and multilayer perceptron; under the Trees file, J48 and random forest algorithms.

2.3.1. Bayesian Classifiers. Bayesian network: the Bayesian network—also known as the belief network—is a probabilistic graphical model that represents knowledge concerning a set of random variables [16]. In this model, each node in the graph represents a random variable, and the edges between the variables represent the conditional dependencies [17]. The conditional dependencies are calculated by statistical probabilistic theories and computational methods. In WEKA software, the BayesNet algorithm is part of the Bayesian file. In order to implement the BayesNet algorithm, the dataset being studied should not have any missing data and all variables must be discrete. In cases where the dataset being studied contains continuous variables as well as discrete variables, discretization can be applied by using the preprocessing tab in the WEKA program. After the discretization is applied to the continuous variables (income and age), the dataset is ready to be studied.

Naive Bayes: the Naive Bayes algorithm is based on the Bayesian theorem and operates on conditional probability. Despite its simplicity, it is a powerful algorithm for predictive modeling. Additionally, the Naive Bayes classifier works quite well concerning real-world situations. An example is spam filtering, which is a well-known problem for which the Naive Bayes classifier is suitable. As with the BayesNet algorithm, there should be no missing data in this algorithm and the variables must be discrete. Since there are no missing data in this dataset, the Naive Bayes algorithm can be applied after discretization of the continuous variables (income and age).

2.3.2. Functions. Logistic regression: logistic regression measures the relationship between a response variable and independent variables, like linear regression, and belongs to the family of exponential classifiers [18]. Logistic regression classifies an observation into one of two classes [19], and this algorithm analysis can be used when the variables are nominal or binary. The data are analyzed after the discretization process for the continuous variables, similar to the Bayesian group.

Multilayer perceptron: the multilayer perceptron algorithm is an artificial neural network algorithm. Artificial neural networks gather information from a training set by minimizing the error iteratively and then applying this information to a new dataset.

2.3.3. Decision Trees. J48 algorithm: this algorithm's name is derived from its tree-like structure and is based on supervised learning techniques. It is a frequently used algorithm due to its ease of implementation, low cost, and reliability. Decision trees' roots consist of decision nodes, branches, and leaves [20]. In the WEKA software, the J48 algorithm uses the rules of the C4.5 algorithm. Therefore, in WEKA, the J48 algorithm is considered a C4.5 algorithm. The C4.5 algorithm can manage numerical values, large data quantities, and datasets with missing values. The C4.5 algorithm uses a threshold value to divide the data into two ranges. The threshold value is selected to provide the most information from the raw data and is determined by sorting the attributes and selecting the average value of the attributes.

Random forest: in this algorithm, the classification process uses more than one "tree" [21]. Each tree produces a classifier, and these classifiers vote to determine the algorithm that gets the most votes [22]. This classification algorithm is then used to classify the dataset.

2.4. Analysis of Classification Algorithms. In DM, it is crucial to use a comparison to determine the best classifier [23]. The classifier's performance is evaluated according to the following criteria [24]:

(i) Classification accuracy: the ability of the model to correctly predict the label of class which is expressed as a percentage

(ii) Speed: the speed refers to the time taken to set up the model

(iii) Robustness: the ability to predict the model correctly even though the data has noisy observations and missing values

(iv) Scalability: the ability of a model to be accurate and productive while handling an increasing amount of data

(v) Interpretability: the level of understanding provided by the model

(vi) Rule Structure: the understandability of the algorithms' rule structure

The WEKA 3.9 software utilizes approximately 100 classification algorithms, and a pre-elimination was carried out by testing the planned data for suitability under various conditions (the algorithms chosen are those that can function in categorical, numerical, binary, or mixed systems). Next, a second screening was carried out that considered criteria such as kappa statistics, the speed of generating the model, usage frequency of the algorithm in the literature, and intelligibility. These eliminations determined that the Bayesian networks, Naive Bayes, J48, random forest, logistic regression, and multilayer perceptron classification algorithms are the most suitable algorithms for our dataset. The next step was comparing the six classification algorithms according to the six statistical criteria (accuracy rate, RMSE, precision, recall, ROC area, and F-measure). The six statistical criteria are explained as follows:

The values of the statistical criteria that are compared to classification algorithms are calculated by using a confusion matrix. The confusion matrix is shown in Table 2.

Accuracy rate (AC): the percentage of correct predictions. According to the confusion matrix, it can be calculated as

$$\mathrm{AC} = \frac{\mathrm{TN} + \mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}, \tag{1}$$

where TN is the true negative, TP is the true positive, FP is the false positive, and FN: false negative.

Precision (P): the fraction of correctly predicted positive observations among the total predicted positive observations.

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{2}$$

where TP is the true positive and FP: false positive.

Recall (R): the fraction of correctly predicted positive observations among all the observations in the class.

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{3}$$

where TP is the true positive and FN: false negative.

F-measure: The Precision and Recall criteria can be interpreted together rather than individually. To accomplish this, we consider the F-Measure values generated by the harmonic mean of the Precision and Recall columns, as the harmonic mean provides the average of two separate factors produced per unit. Therefore, F provides both the level of accuracy of the classification and how robust (less data loss) it is:

$$F - \mathrm{measure} = \frac{2 \times P \times R}{P + R}, \tag{4}$$

where P is the precision and R is the recall.

ROC area: the ROC field curve determines the predictive performance of the different classification algorithms. The area under the ROC curve is one of the essential evaluation criteria used to select the best classification algorithm. When the area under the curve is approaching 1, it indicates that the classification was carried out correctly.

RMSE: the root mean squared error deviation is obtained by determining the square root of the mean squared error (MSE). Normally, the RMSE is used as a measure of the difference between the actual values and the estimated values of a model or estimator. In other words, the RMSE shows the standard deviation of the difference between the estimated

Table 2: Confusion matrix.

		Precision class	
		A	b
Actual class	a	TP	FN
	b	FP	TN

values and the observed values. It is preferable that the RMSE value is small.

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}}. \quad (5)$$

After the analysis using the WEKA data-mining implementation, the classification algorithms were summarized under the selected statistical criteria, as shown in Table 3.

Observing the accuracy percentage in the table, it is clear that the logistic regression algorithm has the highest accuracy rate, while the multilayer perceptron algorithm has the lowest accuracy rate. Similarly, comparing the RMSEs, logistic regression is the best algorithm and multilayer perceptron is the worst with a value of 0.383. The ROC area shows that the logistic regression algorithm's value is the highest, making it the best algorithm. Comparing the values in F-measure according to the recall criterion, the logistic regression algorithm, along with Naive Bayes and BayesNet, all show a good value of 0.824. The algorithms that yielded the best results for the precision criterion are Naive Bayes and BayesNet. In summary, logistic regression is the best algorithm, referring to the columns RMSE, ROC area, accuracy, F-measure, and recall statistical criterions. The only exception is the precision value.

Considering the value of precision, it is clear that the logistic regression algorithm has the closest precision value with the best result of 0.002, which—in the case of this study—does not dramatically affect the results.

2.5. Determination of the Variables Causing Default Risk. The above analysis determined that logistic regression is the best classification algorithm. According to this result, we applied the logistic regression algorithm again to determine the variables that could cause default risk.

The chi-square analysis was applied via the WEKA Select Attribute panel to determine the variables that explain the model the best. Chi-square is an analysis that shows the value of the selected variable depending on the class variable used when the variables are nominal. The analysis recommends subtracting the lowest rank variant from the model.

However, since WEKA cannot apply the chi-square analysis based on the algorithm, it is preferable to exclude variables from the model. The rank results of applying the chi-square analysis to the variables are shown in Table 4.

Studying this analysis, we observe that the age variable is numerically ranked the lowest. Therefore, this variable must either be omitted or converted to a nominal value if it is to be included in the analysis. We therefore transformed the age variable into a categorical variable (one of 3 categories), and

Table 3: Comparison of classification algorithms.

Algorithms/ Properties	Accuracy	RMSE	ROC area	Precision	Recall	F-measure
BayesNet	82.528	0.357	0.836	0.824	0.825	0.824
Naive Bayes	82.532	0.357	0.836	0.824	0.825	0.824
Logistic	83.108	0.342	0.843	0.822	0.831	0.824
J48	82.470	0.367	0.768	0.818	0.825	0.821
Random forest	82.110	0.352	0.828	0.809	0.821	0.814
Multilayer perceptron	81.416	0.383	0.799	0.810	0.814	0.810

Table 4: Chi-square results.

Rank	Attributes
5532.421	10 bills
1846.604	9 house_loan
370.45	2 work
257.218	8 house
30.438	1 gender
10.663	6 health
9.499	7 region
7.173	11 individual_revenue
1.26	4 marriage
0.891	5 education
0	3 age
Selected variables	10, 9, 2, 8, 1, 6, 7, 11, 4, 5, 3 : 11

the logistic regression analysis and the chi-square analysis for variable selection were repeated.

The results of the logistic regression analysis with the nominal age variable are shown in Table 5.

The chi-square analysis with the nominal age variable is shown in Table 6.

The results show that even though the age variable was made nominal, it still ranked the lowest. The logistic regression analysis results with the age variable excluded is shown in Table 7.

Since it is clear from these results that the classification of the data is improved when the age variable is removed, the variable will be removed and the analysis continued.

3. Results and Discussion (Age Variable Excluded)

A logistic regression analysis was performed using 20275 observations and 12 variables (gender, work, marital status, education, health, region, house, home loan, bills, individual revenue, class) as a dataset.

The class subcategory was interpreted according to the non-default status, and a 10-fold cross validation was applied in analysis. The lowest and highest odds ratios are shown in Table 8.

These ratios were interpreted and attributed to different classes without going into default sub groups. According to Table 8, the model predicts that the odds of not going into default risk are 1.1174 times higher for women than they are

Table 5: Nominal age attribute logistic regression results.

Accuracy	82.2491%
Root mean squared error	0.3456
ROC area	0.826
Precision	0.818
Recall	0.822
F-measure	0.820

Table 6: Nominal age attribute chi-squared analysis results.

Rank	Attributes
5532.421	10 bills
1846.604	9 house_loan
370.45	2 work
257.218	8 house
30.438	1 gender
10.663	6 health
9.499	7 region
7.173	11 individual_revenue
1.26	4 marriage
0.891	5 education
0.511	3 age
Selected variables	10, 9, 2, 8, 1, 6, 7, 11, 4, 5, 3 : 11

Table 7: Nominal age attribute omitted logistic regression results.

Accuracy	82.2984%
Root mean squared error	0.3455
ROC area	0.826
Precision	0.819
Recall	0.823
F-measure	0.821

Table 8: Odds ratios per attribute.

Attribute name	Subgroup	Odds ratios
Gender	Women	1.1174
Work	Retired	1.2381
	Looking for a job	0.7632
Marital status	Other	1.0274
Education	Illiterate	1.0453
	Secondary school	0.9666
Health	Good	1.0533
	Medium	0.9184
Region	Mediterranean	1.1063
	Southeastern Anatolia	0.91
House status	Not paying rent	1.1008
Home loan	Yes	0.3449
Bills	Yes	0.0838
Individual revenue	Low income	1.0316
	Medium income	0.9682

for men. This result therefore predicts that women pay their debts on time and tend to default on debt less frequently than men [25].

Concerning the marital status attribute, the model predicts that the odds of not going into default risk are 1.0274 times higher for unmarried individuals than for married individuals. This result therefore predicts that married individuals tend to default on paying their loans more frequently.

The model predicts that the odds of not going into default risk are 1.2381 times higher for a retired person than for people of other work groups. Additionally, people from the "looking for a job" group are 0.7632 times more at risk of going into default than other working groups.

The model further illustrates that the odds of illiterate persons not going into default risk are 1.0453 time higher than that of other education levels. Individuals that have a secondary school degree are most at risk of going into default.

The odds of not going into the default risk are 1.0533 times higher for people in good health than individuals not of good health.

The odds of not going into default risk for people living in the Mediterranean Region are 1.1063 times greater than that of people from other regions. Additionally, those from the Southeastern Anatolia Region are at greater risk of going into default.

The odds for people that are not renting a house to not go into default risk is 1.1008 times higher than that of people renting a house.

The model predicts that the odds of not going into the default risk are 0.3449 for individuals not paying house rent. This means that individuals who do not pay their house rent are more likely to not go into default.

The model predicts that the odds of not going into the default risk are 0.0838 for the nonpayment of bills. This means that individuals who do not pay their bills fully and on time are likely to not pay their other creditors either.

Lastly, the model predicts that the odds of not going into default risk are 1.0316 times higher for people with a lower income level than for others. This indicates that banks should extend lower credit amounts to persons of lower income.

4. Conclusion

Big data analysis cannot be carried out by traditional methods, so data mining is used to extract information from massive amounts of data. DM applications are used intensively in the financial industry for predicting the likelihood of customer default risk.

Our study used an analysis to discover the most suitable classification algorithm to identify credit risks and estimate the likelihood of default. This analysis was carried out using WEKA software and by applying 12 variables such as demographic characteristics of heads of household, total income, debt payment status, and regional information. Six classification algorithms were used (Bayes network, Naive Bayes, J48, random forest, multilayer perceptron, and logistic regression). The performances of the algorithms were compared according to accuracy, root mean squared error, ROC area, F-measure, precision, and recall criteria, and the logistic regression classification algorithm was found to be the best algorithm.

Logistic regression was then applied again to raw data from TUIK by using WEKA 3.9 software to investigate the factors affecting the default risk of individuals using

socioeconomic and demographic variables. Next, a chi-squared analysis was used for attribute selection, which demonstrated that the age attribute needed to be omitted from data. After both these analyses were run, the odds ratio values were used to determine the probability with which individuals with certain characteristics may default on paying loans. These results are not only beneficial to the literature, but they could also have a significant influence in financial institutions to predict the customer default risks.

According to the results, it is observed that women are more likely to honor their payments than men. In the light of the results was obtained, Cigsar and Unal [25] examined the gender variable and investigated the reasons why women were more likely to honor their payments than men. Besides, further research of these findings will be made in the future study.

Regarding the marital status attribute, the results showed that unmarried people have a lower risk of default. The reason for this may be related to the responsibility of married individuals and their possible difficulties in paying their debts on time.

The results for working status show that retirees regularly pay their debts compared to the job-seeking and other inactive groups, which makes the likelihood of defaulting lower. Furthermore, the results for noneducated individuals indicate that their risk of default is lower than that of other education levels. This shows that an increase in the level of education also increases the likelihood of default.

The health status of the individual is also influential in the case of a default, with the results confirming that people in good health are less likely to default than other groups. When considering the region attribute, we see that people from the Mediterranean region are less likely to default. It was also observed that individuals who do not pay house rent were more successful in paying their debts than the ones who do. In this case, we can surmise that the extra responsibility of the rent that the individuals pay makes debt payments difficult. Finally, it was also observed that individuals who do not pay their other debt are likely to default on their loans.

Considering these results and the fact that credit institutions should consider the characteristics of their customers and their circumstances, which will in turn affect their defaults, the risk of default could be reduced by using DM. The risks that could be detected using this method can be eliminated by taking precautionary measures, which could indirectly increase the national income.

This study contributes to existing literature by suggesting classification algorithms that can be used to determine credit risks. Additionally, we identified the variables that can be used in determining the default risk, which will assist future researchers in this field. The study is also valuable in terms of illustrating that DM can be used in the determination of credit risk within the framework of the development of academic studies both in Turkey and globally. Because there are a limited number of studies on the subject of default risk being analyzed using DM applications and WEKA software, this study will contribute to filling the gap in the field.

Lastly, this study proposes a solution for financial institutions and companies in related fields who want to determine credit risks. Future technological development may produce new software and new algorithms for the process. These new algorithms will eliminate the imperfections of existing algorithms and introduce new approaches.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by Cukurova University Scientific Research Fund (Project Number: FYL-2017-8454). The first stages of this study are presented in the International Conference on Applied Analysis and Mathematical Modeling, at Istanbul in July 2017.

References

[1] R. Arora and S. Suman, "Comparative analysis of classification algorithms on different datasets using WEKA," *International Journal of Computer Applications*, vol. 54, no. 13, pp. 21–25, 2012.

[2] J. Xia, F. Xie, Y. Zhang, and C. Caulfield, "Artificial intelligence and data mining: algorithms and applications," *Abstract and Applied Analysis*, vol. 2013, Article ID 524720, 2 pages, 2013.

[3] W. H. Inmon, *Building Data Warehouse*, QED/Wiley, Hoboken, NJ, USA, 2005.

[4] I. O. A. In Belgium, *Big Data: An actuarial perspective*, Institute of Actuaries, Belgium, 2015.

[5] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, Burlington, MA, USA, 2012.

[6] D. Donko and A. Dzelihodzic, "Data mining techniques for credit risk assessment task," *Recent Advances in Computer Science and Applications*, pp. 105–110, 2013.

[7] S. Belkıs, *Credit Rating*, Capital Market Licensing Registration and Training Organization, Turkey, 2016.

[8] H. Selimler, "Analysis of problem loans, assessment of the effect on bank financial statements and rates," *Journal of Financial Research and Studies*, vol. 7, no. 12, 2015.

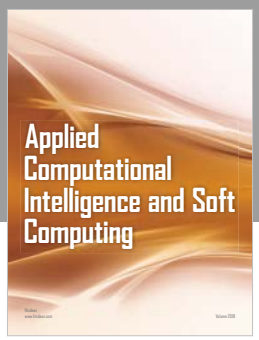
[9] T. V. Gestel and B. Baesens, *Credit Risk Management: Basic Concepts: Financial Risk Components, Rating Analysis, Models, Economic and Regulatory Capital*, Oxford University Press, Oxford, UK, 2008.

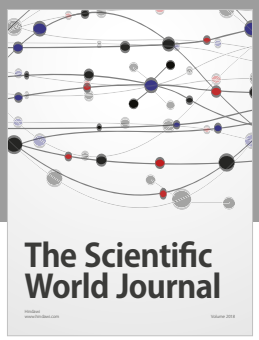
[10] A. Wang, L. Yong, W. Zeng, and Y. Wang, "The optimal analysis of default probability for a credit risk model," *Abstract and Applied Analysis*, vol. 2013, Article ID 878306, 9 pages, 2014.

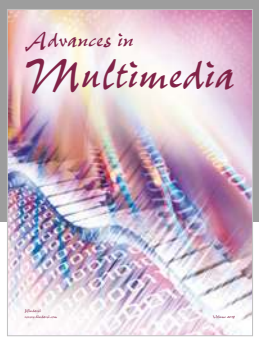
[11] C. Huang, M. Chen, and C. Wang, *Credit Scoring a Data Mining approach based on support vector machines*, Vol. 33, New York City, NY, USA, 2016.

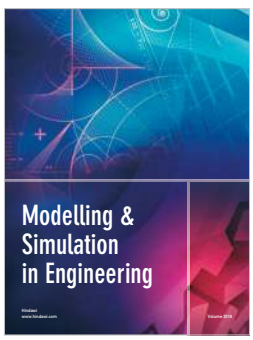
[12] G. Kou and W. Wu, "An analytic hierarchy model for classification algorithms selection in credit risk analysis," *Mathematical Problems in Engineering*, vol. 2014, Article ID 297563, 7 pages, 2014.

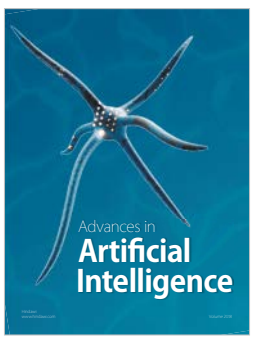
[13] M. D. M. Sousa and R. S. Figueiredo, "Credit analysis using data mining: application in the case of a credit union," *Journal of Information Systems and Technology Management*, vol. 11, no. 2, pp. 379–396, 2014.
[14] I. Yeh and C. Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," *Expert Systems with Applications*, vol. 36, no. 2, pp. 2473–2480, 2009.
[15] G. Kesavaraj and S. Sukumaran, *A Study on Classification Techniques in Data Mining*, IEEE, 2013.
[16] Y. Yan and B. Suo, "Risks analysis of logistics financial business based on evidential bayesian network," *Mathematical Problems in Engineering*, vol. 2013, Article ID 785218, 8 pages, 2013.
[17] I. B. Gal, F. Ruggeri, F. Faltin, and R. Kenett, *"Bayesian networks," Encyclopedia of Statistics in Quality and Reliability*, Wiley and Sons, Hoboken, NJ, USA, 1st edition, 2007.
[18] S. Chen, Y.-J. J. Goo, and Z.-D. Shen, "A hybrid approach of stepwise regression, logistic regression, support vector machine, and decision tree for forecasting fraudulent financial statements," *The Scientific World Journal*, vol. 2014, Article ID 968712, 9 pages, 2014.
[19] Y. Kumar and G. Sahoo, "Analysis of parametric and non parametric classifiers for classification technique using WEKA," *International Journal of Information Technology and Computer Science*, vol. 4, pp. 43–49, 2012.
[20] M. Kantardzic, *Data Mining: Concepts, Models, Methods, and Algorithms*, John Wiley & Sons, Danvers, MA, USA, 2003.
[21] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[22] T.-T. Nguyen and J. Z. Huang, "Unbiased feature selection in learning random forests for high-dimensional data," *Scientific World Journal*, vol. 2015, Article ID 471371, 18 pages, 2015.
[23] I. Nguyen, E. Frank, and M. Hall, *Data Mining Practical Machine Learning Tools and techniques*, Morgan Kaufmann, Burlington, MA, USA, 2011.
[24] J. Stefanowski, *Data Mining- Evaluation of Classifiers*, Institute of Computing Sciences Poznan University of Technology, Poland, 2010.
[25] B. Cigsar and D. Unal, "The effect of gender and gender-dependent factors on the default risk," *Revista de Cercetare si Interventie Sociala*, vol. 63, pp. 28–418, 2018.

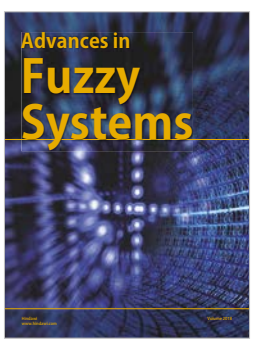












Hindawi
Submit your manuscripts at
www.hindawi.com