Journal of
Computational and Theoretical Nanoscience
Vol. 16, 3489–3503, 2019

Loan Default Prediction Model Using Sample, Explore, Modify, Model, and Assess (SEMMA)

Hafiz Ilyas Tariq, Asim Sohail, Uzair Aslam, and Nowshath Kadhar Batcha*

School of Computing, Asia Pacific University of Innovation and Technology, 63000, Malaysia

The purpose of this study is to provide a comprehensive research and to develop a model to predict the loan defaults. This kind of models becomes inevitable as the issue of bad loans are very much critical in the financial sector especially in micro financing banks of various underdeveloped and developed countries. To cope up with this problem a comprehensive literature review was done to study the significant factors that leads to this issue. Moreover, these reviewed studies were critically focused towards applying data mining techniques for the prediction and classification of the loan defaults. This study used methodologies named KDD, CRISP-DM and SEMMA. While in the experimentation phase, three different data mining techniques were applied for the proposed model and their performances were evaluated on various parameters. Based on these parameters, the best method was selected, explained and suggested because of its significant characteristics regarding the prediction of the loan defaults in the financial sector.

Keywords: Data Mining, Loan Defaults Loan Lending, SEMMA.

RESEARCH ARTICLE

1. INTRODUCTION

Without a doubt, financial lending services hold a great amount of significance for any individual, business or enterprise. As such services are required by an individual or a business to achieve or accomplish their goals and to compete with the giants of their fields [1]. Financial loans are a major part of the primary source of capital not only in the emerging economies but also in the developed capital markets by both individuals and enterprises. As the lending growth by the financial firms and the banks are considered as the key factor for inflation level and interest rate of any country which drives its economic growth and depicts its economic condition. According to the mission statement of the study in the role of financial services, the economic growth of the real economy is the primary role of the financial firms [2]. With such great importance and benefits of financial lending comes some major issues and bottleneck problems. The most common and substantial issue in the domain of financial lending is the fair and successful lending of loans while keeping the ratio of loan defaulters to the least minimal value. In the financial lending, the risk of loan defaulters can never be neutralized but can be minimized [3].

According to the study, during the year of 2017 in India, only the bad loans crossed the threshold of about 207 billion dollars equaling the percentage of about 9.6 for loan defaulter ratio. While the greatest number of loan defaulter cases were registered in Italy making the total of 16.4% for the loan defaulters [4]. To cope up with the issue of high ratio loan defaulters' lot of work has been done.

As in the past couple of decades, the decision making for the financial lending has been very much influenced by the information sharing and technological advancements. While employing various credit scoring models that include FICO Scoring Model, Vantage Score Model, Credit Xpert Credit Score [5].

The technique of credit scoring is to evaluate different credit attributes by analyzation and classification to an individual and enterprise profile to assess the credit decision or to estimate the creditworthiness [6].

Only credit scoring is not sufficient for the financial lending because of such a massive number of loan defaulters. Some financial expert's judgmental reviews have to be accompanied subjectively. As financial analysts not only rely on the credit scores but also on their experience regarding the historical successful and unsuccessful cases as well for better decision making [7]. Moreover, with such tremendous growth of the financial lending and to improve the credit defaulter ratio, advanced statistical methods were introduced to fill the gap of underperforming credit scoring models. These advanced statistical models such as genetic programming and neural networks provided the alternative from the previous traditional statistical models which were based on the logistic regression and discriminant analysis [8].

*Author to whom correspondence should be addressed.

In the more recent years, different researchers have also employed different data mining techniques for the loan defaulter predictions. The authors have proposed a study regarding the comparison of an advanced classification algorithm for credit scoring against the financial lending request [9]. The main objective of this research is to study the previous work regarding the field of financial lending and evaluate the different approaches which have been used for the prediction of loan defaulters. Aiming to the objective, the purpose of this study is to employ various data mining techniques on the collected dataset from Kaggle repository to predict the successful financial lending [10]. Finally, the proposed predicted model will be evaluated by benchmarking its performance against other modeling techniques.

The rest of this study is systematized as follows. Chapter 2 discusses the researcher work of various authors from the field of financial services while focusing only on financial lending decision making. Chapter 3 focuses on the approach employed in this study for the modeling purpose. Chapter 4 describes the experimentation phase and discusses the interpretation of the results as well. Chapter 5 will summarize the entire work of the proposed study and present the conclusion regarding the projected objective.

2. RELATED WORKS

In the last couple of decades, credit scoring techniques has received much attention in the literature and the objective of these studies is to provide a competitive edge for the retail banks in such a cut-throat competition of core banking [11]. As nowadays due to the availability of big data, different data mining techniques have been becoming very popular to gather useful insights from such amount of data. Similar approaches have been adopted by multiple researchers in the credit scoring field who have tried to assess the financial lending risk by employing different data mining technique [12]. In the banking sector, a huge amount of data is being generated and collected on a daily basis in the form of collateral details, transaction details, credit card details, risk profiles, loans and customer information. Moreover, a number of decisions in the banking environment is purely based on this critical information [13]. All those researchers have tried to incorporate these attributes to gain some knowledgeable insights for financial lending assessment to reduce the credit risk and to help the retail banking system. As such, this following section will analyze few traditional data mining techniques for credit scoring purpose.

Decision making in the field of machine learning is the most valuable and first most objective. Therefore, the decision tree classification technique is very much appropriate to have such decision making in the data analysis which can be explicitly and visually represent [14].

In another study the authors have proposed the prediction of the loan defaulters by including the relationship of borrower mobile phone usage with the other loan default variables. Three different variables from the phone data were selected based on the high significance of the loan defaulter. The variables names are mobility patterns, telecommunication patterns, and App usage patterns. These variables are extracted carefully by using the recursive feature elimination (RFE) on the real data set. To keep the privacy of the individuals the contact details such as the name, ID, and phone numbers were encrypted. With on the selected variables AdaBoost algorithm was applied as a decision tree classifier [15].

While another study came up with the different approach of CRISP-DM (Cross Industry Standard Process for Data Mining) framework. In their study for the 1st step they considered was the banking process to approve the loan and to understand the 5 C's used by baking officers to approve or reject the loan. These C's are the cash flow, conditions of the borrower, collateral, credit history, capitalization, and conditions. The dataset from the banking industry was collected by keeping the 5 C's in the mind. In the dataset out of 8 variables, 6 are nominal and 2 are a numeric type. The data set used was in Attribute-Relation file format (ARFF) which is accepted by the Weka program. Three different models were formed which are j48, Naïve Bayes and Bayes Net [16].

To improve the accuracy of decision tree technique and to make them competitive and interactive with other classification techniques some sort of bagging for the construction of multiple trees or boosting for the purpose of iterative learning is required. To fulfill this bagging requirement, a random forest is very much useful as it forms a forest of random decision trees [17]. For the credit loan prediction, models based on the random forest are equipped with much more booting and baggage characteristics [18]. In this section, a study related to credit scoring with the employment of random forest technique has been discussed.

Whereas, the authors have proposed a study for the analysis of credit default risk by the implementation of machine learning algorithm. Credit score is not only very much crucial to customers but also for the financial firms and banks to evaluate the request for the financial lending accordingly. In the proposed study, the authors have implemented various data mining classification techniques to predict the financial loan defaulters to maximize the successful financial loan lending ratio. The target variable i.e., bad credit had the most impact because of the duration of the credit, amount of the credit and the age of the creditor and these features were extracted by the feature engineering. Furthermore, the authors have implemented different data mining classification techniques that include Random Forest, Adaptive Boosting, Support Vector Machine, Linear Regression and Neural Network for the credit default prediction based on binary classification. The results of the study showed that the Random Forest

classifier outperformed all the other classification models because of its highest accuracy [19].

Among many classifiers used for the prediction of the loan default, the researchers have proposed a prediction model. Their study had 2 main objectives, one is to find and explore the hidden information and relations among the attributes and to accurately predict the loan defaulters by using the explored variables. For the best classification, the results of 7 different classifiers were compared and the best one was selected. After extensive experimentation on all the 7 classification algorithms, the best algorithm that outperformed the rest of the others was the Random forest with the highest accuracy [20].

Bhargava in his study have compared different techniques for the identification of loan defaulters. This model gives the probability of loan default based on the customer characteristics. The dataset for their research was obtained from an online marketplace called Lending Club. 1-R Square (a statistical technique) was used for the selection of variables used based on the correlation and significance. Same data was used in 4 models: Decision tree, Logistic Regression, Random Forest and Neural Network. The results of all the models are calculated in terms of the minimum misclassification rate. From the results, the minimum misclassification was of the random forest model, which performed fairly better than the rest of the models [21].

Credit scoring is considered the core element for detecting and avoiding any financial risk. The authors have proposed some advanced data mining models to predict the loan default customers by keeping the credit score as the main component. To improve the loan approval process and to predict the defaulter borrower, the data set is obtained from the microfinance institution. The dataset comprised upon client information which include employment, credit history, financial, demography and behavioral attributes. For the implementation of the models, the data is preprocessed by using the Oracle Data Miner (ODM). For the experimentation, generalized linear model is used for the prediction which is an improved form of regression that can also predict the confidence span where the probability is positive [22].

Neural networks are considered as the most advanced techniques in the field of data mining. Their system is based on the input variables as the explanatory variables which are connected to the output variables by a number of hidden layers for the purpose of interactions [23]. Many authors have used this technique for the prediction purposes in the financial sector to improve a number of services provided to the customers and this section have discussed few of these approaches in this section.

A novel approach has been presented in the study for credit scoring using employing credit default swaps while applying deep learning technique. As of these days, the major decision making regarding the credit risk management is based on the corporate credit scoring [24]. The other key aspect of the proposed study is related to the credit default swaps (CDS) due to the credit risk exposures and financial crises. This study has tried to comprehend the gap in the domain of financial lending by employing a deep learning model known as deep belief network (DBN) model with the credit rating based on the data particular to the credit default swaps. The incorporation of deep belief network (DBN) for credit scoring would improve the processing for efficient credit scoring in particular to the CDS market. To carry out the proposed study, the authors have collected the CDS data. Before the application of deep machine learning algorithm by Restricted Boltzmann Machines, data pre-processing has been employed over the collected dataset to enhance the performance of the deep belief network (DBN). The results of the study showed that all the algorithms which were employed for the validation purpose were outperformed by the deep belief network (DBN). Furthermore, authors have suggested that with much more rich information regarding the credit lending could improve the performance of the model and would make it more applicable in the financial industry.

3. EXPERIMENTATION AND RESULTS

The process of finding and extracting meaningful more importantly useful information from a collection of data is considered as data mining. As the significant objective of the data mining process is to discover relationships which were previously unknown by employing various multi diversionary skill set based on different statistical techniques, machine learning approaches and methodologies. These data mining methodologies are used as assumptions in the data mining process to achieve reliable and applicable systematic approaches to attain hidden and meaningful information of the highest degree.

The process of finding and extracting meaningful more importantly useful information from a collection of data is considered as data mining. As the significant objective of the data mining process is to discover relationships which were previously unknown by employing various multi diversionary skill set based on different statistical techniques, machine learning approaches and methodologies. These data mining methodologies are used as assumptions in the data mining process to achieve reliable and applicable systematic approaches to attain hidden and meaningful information of the highest degree.

The significant difference between the SEMMA and rest of the two methods which are CRISP-DM and KDD is the requirement and understanding of the pre-requisites of the business understanding and databases. Moreover, even after the implementation phase of the model, both methodologies KDD and CRISP-DM require the model estimation for their performance evaluation which is also not required in SEMMA. All these major critical differences between SEMMA and other data mining methodologies provides a

considerable advantage in the employment of the SEMMA methodology.

3.1. Sample

The first process of SEMMA data mining process is known as sampling and this step of the employed methodology is very much considered as the optional process. As the main objective of the sampling phase is to obtain a well-represented sample data from the entire population data. The process of collecting data from the entire population is a significantly difficult task due to such difficulty, SEMMA provides an option of using a sample population data for the development of the model. While in this study, the data sample was obtained from Kaggle Repository [27].

3.2. Explore

The next phase of the SEMMA methodology is related to the review of the data which is also known as exploration. This phase can be considered as one of the most vital steps as in this phase, the understanding related to the structure of the data has to be developed which is very much substantial for model development. All this exploration of the data will improve the discovery process of the meaningful information and also helps in the identification of the anomalies in the data as well.

3.2.1. Dataset Structure

For the development of the proposed model, the provided data is divided into two datasets each for training purpose of the model and for the performance testing of the model. Both of the datasets have 13 attributes related to the feature of individuals although the testing dataset has no values in its class variable as they have to be predicted by the developed model. Furthermore, the attributes are comprised of seven categorical, four continuous, applicant's id and a class variable. Moreover, the training dataset has 614 observations while the testing dataset has 367 observations where each observation represents the financial and

The CONTENTS Procedure

Data Set Name	DAP.SME_TRAIN	Observations	614
Member Type	DATA	Variables	13
Engine	V9	Indexes	0
Created	07/28/2019 07:07:29	Observation Length	152
Last Modified	07/28/2019 07:07:29	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Data Set Name	DAP.SME_TEST	Observations	367
Member Type	DATA	Variables	13
Engine	V9	Indexes	0
Created	07/28/2019 07:09:18	Observation Length	152
Last Modified	07/28/2019 07:09:18	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Fig. 1. Dataset dimensions.

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Informat
9	CANDIDATE_INCOME	Num	8	
6	EMPLOYMENT	Char	14	$14.
4	FAMILY_MEMBERS	Num	8	BEST32.
2	GENDER	Char	14	$14.
10	GUARANTEE_INCOME	Num	8	
11	LOAN_AMOUNT	Num	8	
8	LOAN_APPROVAL_STATUS	Char	14	$14.
12	LOAN_DURATION	Num	8	
13	LOAN_HISTORY	Num	8	
7	LOAN_LOCATION	Char	14	$14.
3	MARITAL_STATUS	Char	14	$14.
5	QUALIFICATION	Char	14	$14.
1	SME_LOAN_ID_NO	Char	14	$14.

Fig. 2. Structure of the data.

Table I. Description of the dataset.

Attribute	Data type	Description
SME_LOAN_ID_NO	Categorical	Reference no for the loan
GENDER	Categorical	Male/Female
FAMILY_MEMBERS	Categorical	Total no of family members
MARITAL_STATUS	Categorical	Married/Not married
EMPLOYMENT	Categorical	Yes/No
QUALIFICATION	Categorical	Graduate/Under graduate
LOAN_AMOUNT	Continuous	Amount applied for in thousands
LOAN_DURATION	Continuous	Repayment duration for the loan
LOAN_HISTORY	Categorical	Past loan records positive/negative
LOAN_LOCATION	Categorical	City/Town/Village
CANDIDATE_INCOME	Continuous	Monthly income of the applicant
GUARANTEE_INCOME	Continuous	Joint applicant income
LOAN_APPROVAL_STATUS	Categorical	Class variable of loan Yes/No

Variable	Mean	Std Dev	Minimum	Maximum	Median	N	N Miss	Std Error	Variance	Mode
CANDIDATE_INCOME	5403.46	6109.04	150.0000000	81000.00	3812.50	614	0	246.5408575	37320390.17	2500.00
GUARANTEE_INCOME	1621.25	2926.25	0	41667.00	1188.50	614	0	118.0937733	8562929.52	0
LOAN_AMOUNT	146.4121622	85.5873252	9.0000000	700.0000000	128.0000000	592	22	3.5176174	7325.19	120.0000000
LOAN_DURATION	342.0000000	65.1204099	12.0000000	480.0000000	360.0000000	600	14	2.6585296	4240.67	360.0000000

Fig. 3. Summary statistics of continuous attributes.

socio-economic features of the respective loan applicant. The dimensions of the training and testing dataset can be observed in the mentioned below Figure 1.

Structure of the data is also very much significant in the exploration process because to perform any modification on the data or its attributes, their datatype and formats have to be sorted out first. For this purpose, Figure 2 can be observed where each attribute is mentioned along with their data types.

Additionally, from Figure 2 it can also be seen that all the variables are either characters or numeric in the type column as SAS only supports these two data types while the length column shows the number of bits required to store each instance of the corresponding variable.

Moreover, the continuous and categorical variables in the provided dataset are mentioned in Table I along with their brief explanation.

3.2.2. Continuous Attributes Exploration

As mentioned in the above table that there are only four continuous variables. In this section, those continuous attributes will be explored by employing different visualizations and other statistical techniques. While Figure 3 shows the summary, statistics related to the all four continuous variables and there are few missing values which are 22 and 14 in "LOAN_AMOUNT" and "LOAN DURATION" respectively.

Moreover, Figure 4 is illustrating the representation of four continuous attributes in terms of histograms with

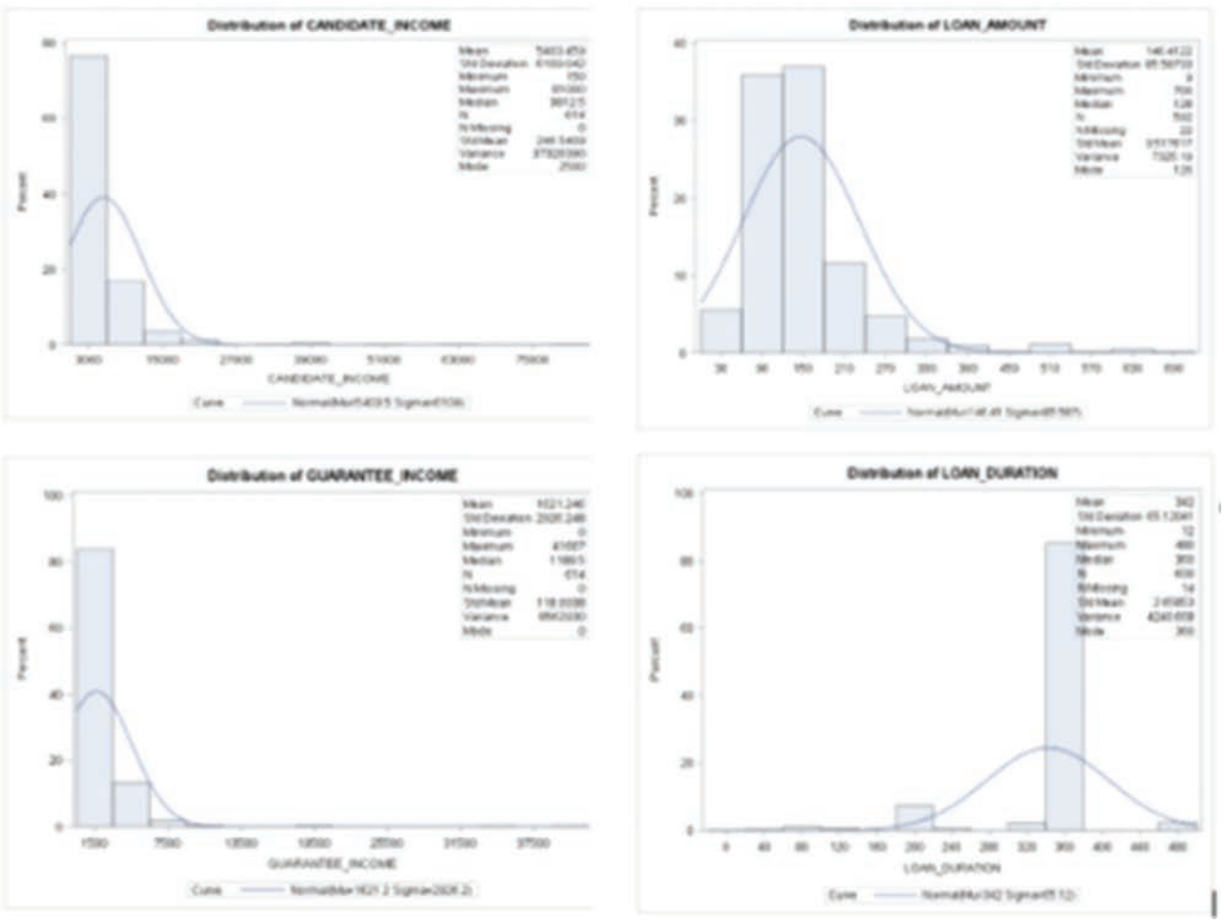


Fig. 4. Histogram related to continuous attributes.

RESEARCH ARTICLE

their density curves. These plots also show that "CANDIDATE_INCOME," "LOAN_AMOUNT" and "GUAREENTEED_INCOME" are very much rightly skewed the other attribute "LOAN_DISTRIBUTION" it represents that majority of the loan applicants have been previously participating in other loan lending services.

Additionally, Figure 5 is representing the distribution of the continuous variables with respect to the class variable. Here, it can be observed that the majority of the sample population belonged to a low-level income which is less than 5000 dollars and their requested loan amount are under 150 thousand dollars. While there are few applicants with very large income although they cannot be considered as the outliers because their "GUAREENTEED_INCOME" is also very high.

Furthermore, in the exploration of the "LOAN_AMOUNT" against "CANDIDATE_INCOME" by employing scatter plot shown in Figure 6, it can be observed that most of the people whose "EMPLOYMENT" status is "NO" are in majority with loan applications in comparison to the applicants who are currently employed.

3.2.3. Categorical Attributes Exploration

Initially, the summary statistics have been collected for the exploration of the seven categorical attributes which represents the number of instances of each categorical attribute with respect to their categories. Moreover, Figure 7 also represents the number of missing values in each attribute along with the percentage of each category in their respective categorical variable.

From this exploration step, it can be identified that the handling of the missing values very much crucial in the next process of the employed SEMMA methodology.

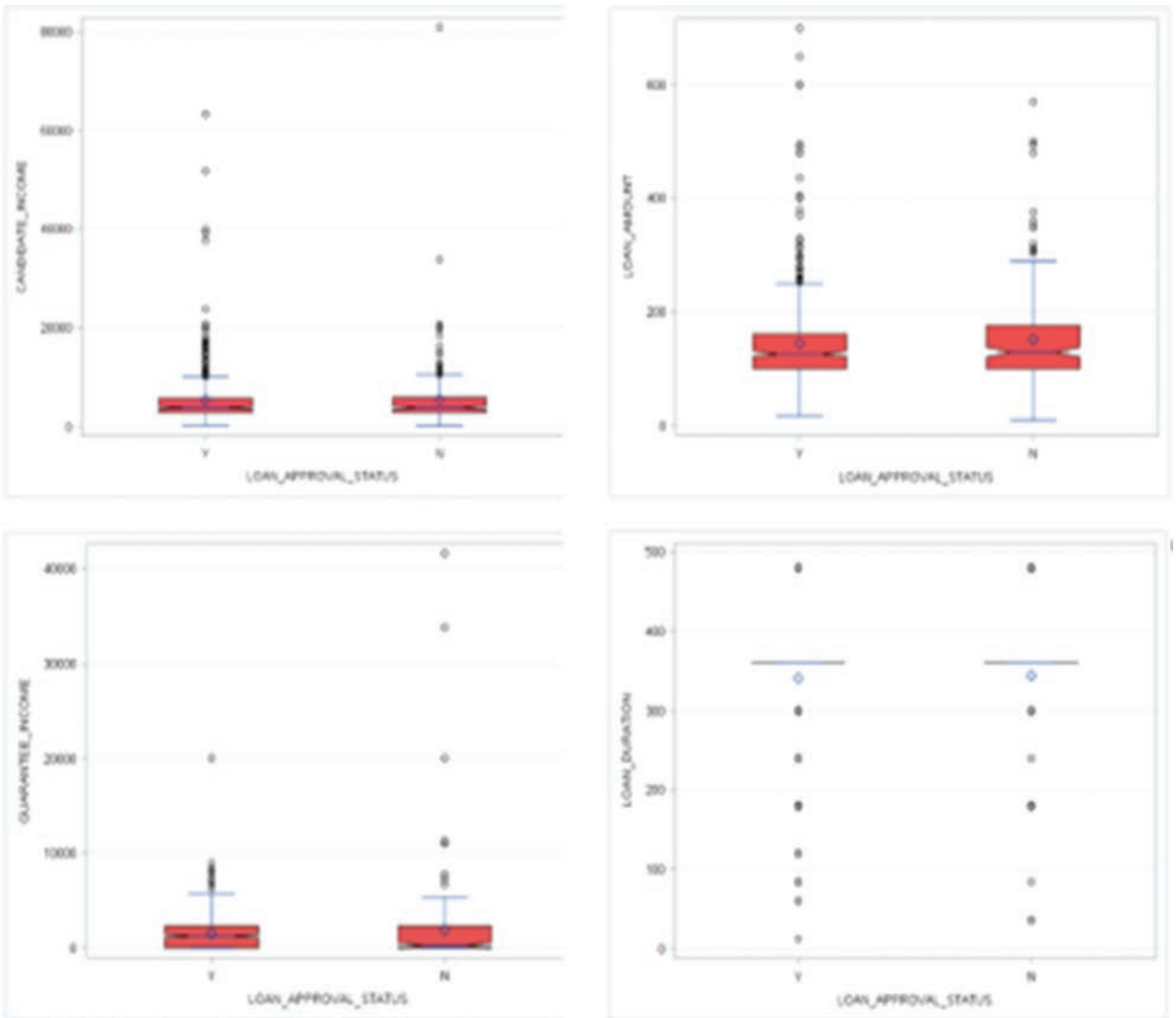


Fig. 5. Box plot related to continuous attributes.

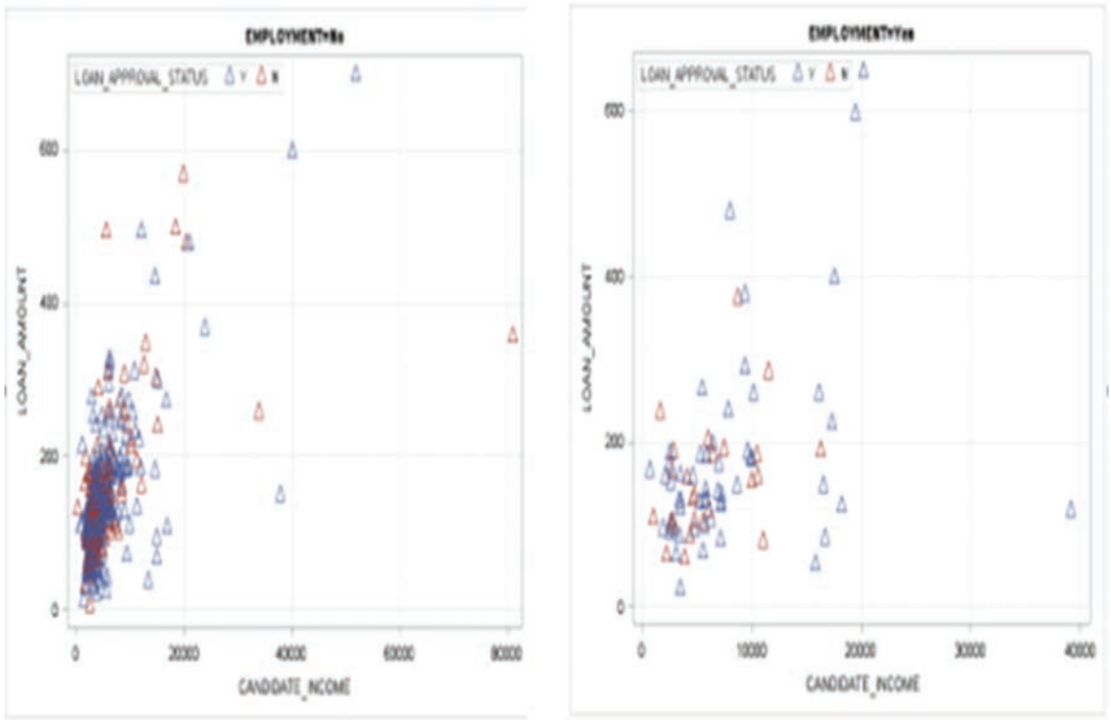


Fig. 6. Scatter plot between income and loan amount against employment status.

GENDER	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Female	112	18.64	112	18.64
Male	489	81.36	601	100.00
Frequency Missing = 13				

LOAN_HISTORY	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	89	15.78	89	15.78
1	475	84.22	564	100.00
Frequency Missing = 50				

MARITAL_STATUS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Married	398	65.14	398	65.14
Not Married	213	34.86	611	100.00
Frequency Missing = 3				

LOAN_LOCATION	Frequency	Percent	Cumulative Frequency	Cumulative Percent
City	202	32.90	202	32.90
Town	233	37.95	435	70.85
Village	179	29.15	614	100.00

EMPLOYMENT	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	500	85.91	500	85.91
Yes	82	14.09	582	100.00
Frequency Missing = 32				

FAMILY_MEMBERS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	345	56.19	345	56.19
1	102	16.61	447	72.80
2	101	16.45	548	89.25
3	66	10.75	614	100.00

QUALIFICATION	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Graduate	480	78.18	480	78.18
Under Graduate	134	21.82	614	100.00

Fig. 7. Summary statistics of categorical attributes.

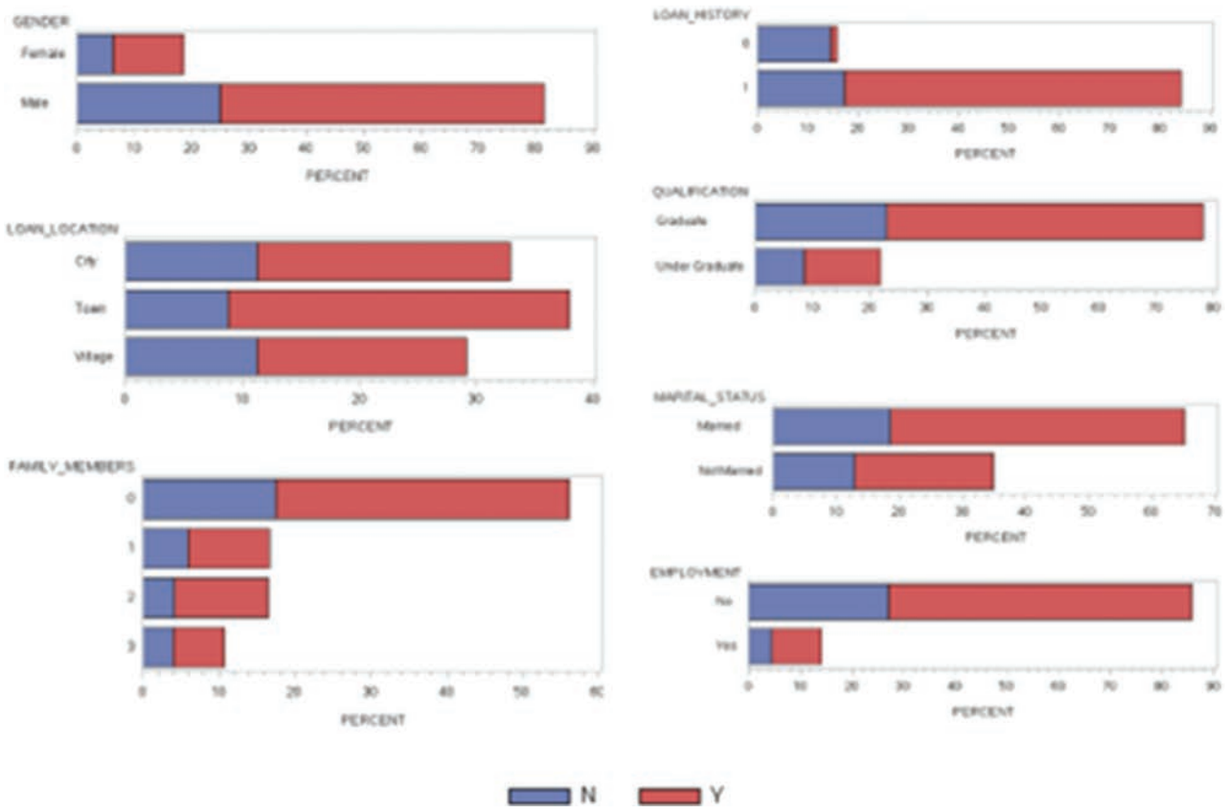


Fig. 8. Distribution of categorical attributes with respect to the class variables.

As there are 13, 50, 3, 32, missing values in "GENDER," "LOAN_HISTORY," "MARITAL_STATUS" and "EMPLOYMENT" respectively which can be observed in Figure 7.

While from Figure 8 related to the exploration of the categorical variables, it can be very much clearly visualized that the majority of the loan applicants were unemployed graduates and they were already part of some loan lending service. Moreover, most of these applicants have no dependents family members to take care of and they were belonging to the male category of "GENDER" attribute as well. Furthermore, the overall representation of the categorical variables depicts that approved loans are in much higher percentage with respect to the rejected ones.

The next phase will explain the modifications required before the implementation of the data mining techniques for loan prediction model development.

3.3. Modify

In SEMMA methodology, the third and last process before the employment of the muddling phase is the modification step. This step has a great significance in the data mining methodology because all the critical data pre-processing techniques have to be implemented in this step.

Major tasks related to the modification of the data in the proposed study are the conversion of the data types and handling of the missing values. As all the categorical attributes with character data type have to be converted into the numeric for the purpose of the algorithm application. Moreover, all the missing values in the entire data has to be handled for improved and efficient modeling.

3.3.1. Data Structure Modification

The first step in the modification task is related to the data structure of the categorical attributes. As all the values in the categorical attributes are in the character format which cannot be incorporated during the implementation of the data mining algorithm. To cope up with this issue all the values in the categorical attributes are converted to into numeric as factors. Once the categorical attribute instances are changed from character to numeric, the next step is to convert the data types of the character attributes to numeric.

3.3.2. Missing Data Modification

The next step in the modification process is related to the handling of the missing values in the dataset. Before the implementation of this task, Figure 9 shows the number of missing values with respect to their attribute.

The MEANS Procedure

Variable	N Miss
FAMILY_MEMBERS	0
CANDIDATE_INCOME	0
GUARANTEE_INCOME	0
LOAN_AMOUNT	22
LOAN_DURATION	14
LOAN_HISTORY	50
GENDER	13
MARITAL_STATUS	3
EMPLOYMENT	32
QUALIFICATION	0
LOAN_LOCATION	0
LOAN_APPROVAL_STATUS	0

Fig. 9. Missing values in the dataset.

Various techniques are available in SAS to handle these missing values although, in this study, these missing values are handled by the "Median" method. As in this method, each missing value is replaced by the median of that particular attribute and this process continues until all the missing values are handled. The implementation of this technique and its results can be observed in Figure 10 where "PROC STDIZE" was employed.

3.4. Model

The fourth vital step after the modification phase is known as modelling. In this step of SEMMA, different data mining algorithms and techniques are employed for the development of the proposed model. The aim of this step is to identify the hidden and meaningful information from the pre-processed dataset. To implement this task for the proposed model development regarding the prediction of the loan lending, three different data mining techniques will be employed, and their employment has been discussed in this section.

3.4.1. Decision Tree (DT)

The first model is developed by employing "Decision Tree" algorithm over the dataset which has been passed through the modification stage. The first step is to build a tree from the attributes of the training dataset to for the purpose of training and in Figure 11, the implemented structure of the decision tree model can be observed which was employed by the "PROC HPSPLIT" function was used.

The model has been designed based on the classification variable "LOAN_APPROVAL STATUS" which has been shown in Figure 12. Moreover, the growth of the tree has been controlled by calculating the homogeneity based on entropy. Furthermore, to reduce the size of the tree pruning function of "COST COMPEXITY" with a maximum number of 12 leaves. While in Figure 15 the first 7 leaves of the tree can be seen where the node splitting was depending upon the most homogeneous attribute "LOAN_HISTORY" and then based on "GUARANTEED_INCOME."

Below mentioned Figure 13 represents the attributes based on their homogeneity calculated by the entropy which has been used in the building of the proposed tree.

The MEANS Procedure

Variable	N Miss
FAMILY_MEMBERS	0
CANDIDATE_INCOME	0
GUARANTEE_INCOME	0
LOAN_AMOUNT	0
LOAN_DURATION	0
LOAN_HISTORY	0
GENDER	0
MARITAL_STATUS	0
EMPLOYMENT	0
QUALIFICATION	0
LOAN_LOCATION	0
LOAN_APPROVAL_STATUS	0

Fig. 10. Handling of the missing values.

Model Information	
Split Criterion Used	Entropy
Pruning Method	Cost-Complexity
Subtree Evaluation Criterion	Number of Leaves
Number of Branches	2
Maximum Tree Depth Requested	10
Maximum Tree Depth Achieved	10
Tree Depth	7
Number of Leaves Before Pruning	93
Number of Leaves After Pruning	12
Model Event Level	0

Number of Observations Read	614
Number of Observations Used	614

Fig. 11. Structure of decision tree model.

RESEARCH ARTICLE

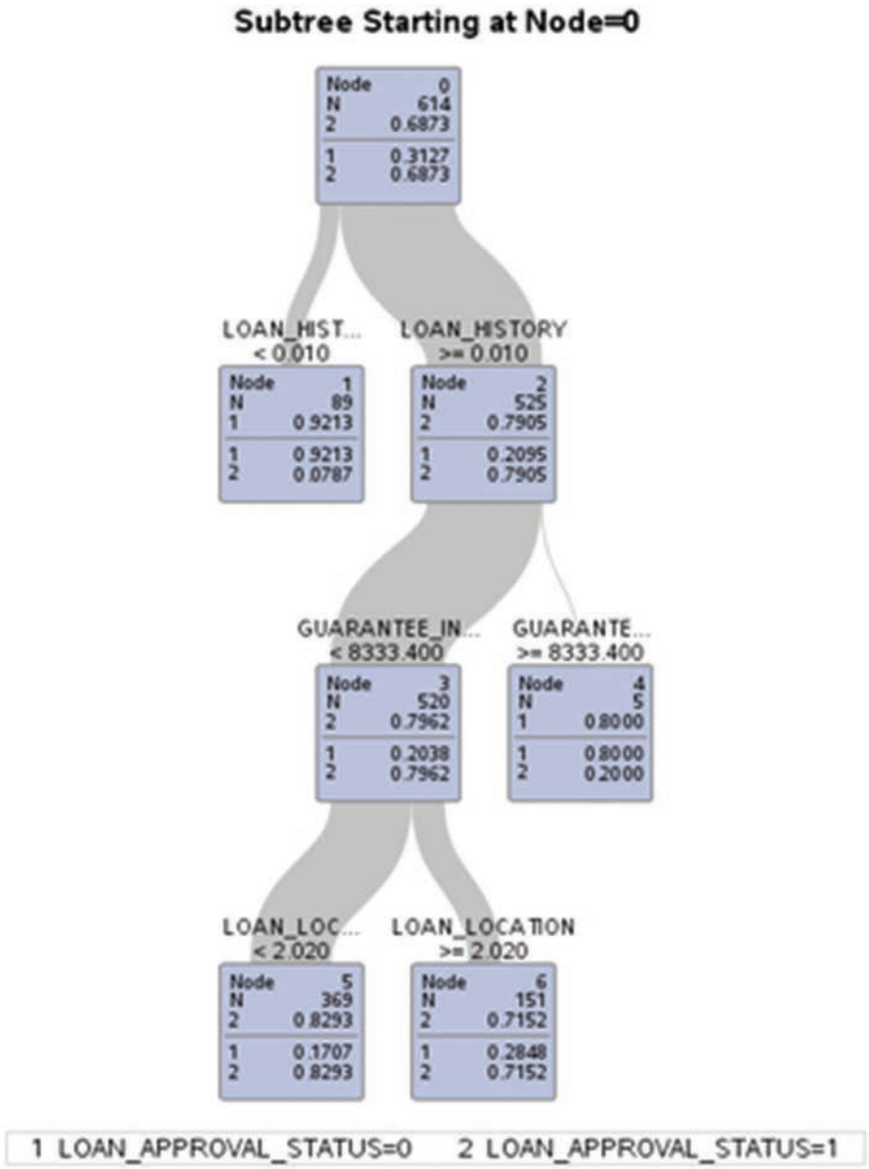


Fig. 12. Decision tree.

This explains the selection criteria related to the splitting of the terminal and internal nodes.

Moreover, the next Figure 14 shows the confusion matrix related to both the training model and validation model. The accuracy achieved in the validation of the model by employing 10-cross-validation technique was 79.8 percent.

The next visualization is related to the receiver operating characteristics (ROC) which is used to evaluate the quality or to evaluate the performance related to the diagnostic tests. Another easy way to understand the ROC plot is by observing the area under the curve (AUC) which is 0.82 in the case of the proposed model. These fact and figures are mentioned in Figure 15.

3.4.2. Logistic Regression (LG)

The second model is developed by employing "logistic regression" algorithm. The modified dataset was used, and this technique was employed. The model was built on the attributes of the training dataset and for the purpose of training. Figure 16 shows the implementation and the structure of the logistic regression model in which "PROC LOGISTIC" function was employed.

This model of logistic regression has employed stepwise regression along with the Fisher test and only those variables will enter the model which are significant by having their $Pr < 0.05$. This criterion was only satisfied by two attributes "LOAN_HISTORY" and

Variable Importance			
	Training		
Variable	Relative	Importance	Count
LOAN_HISTORY	1.0000	8.7817	1
LOAN_AMOUNT	0.3104	2.7258	3
GUARANTEE_INCOME	0.2654	2.3306	2
FAMILY_MEMBERS	0.1926	1.6910	1
LOAN_LOCATION	0.1901	1.6694	1
QUALIFICATION	0.1790	1.5720	1
CANDIDATE_INCOME	0.1740	1.5282	1
MARITAL_STATUS	0.1601	1.4056	1

Fig. 13. Importance of the attributes in decision tree.

"MARITAL_STATUS." Figure 17 shows the significant variables with their *Pr* values.

Moreover, in Figure 18 there are a couple of very crucial parameters which has to be discussed and the first parameter is Percent Concordant. This parameter explains the effective probability threshold from which an observation is classified to either level 1 or 0. The value of this parameter should always be a higher value which means that the percentage regarding the probability of approval is higher than the probability of disapproval. In the proposed case the value of the Percent Concordant parameter is 59.7 percent. The second parameter from this figure which needs to be discussed is the C statistics. It is the classification rate of the of the developed model and its value at least has to be greater than 0.5 for an effective model. As in the case of the C value which is 0.744 shows that the area under the curve (AUC) is 74.4 percent and this result can also be confirmed in the receiver operating characteristics (ROC) curve plot as well.

The next visualization (Fig. 19) is related to the receiver operating characteristics (ROC) curve and area under the curve (AUC) which is 0.74 in case of the proposed model and can be observed in Figure 19. While the accuracy of the model can also be calculated from the correct and incorrect events and non-events which results into 80.9 percent while the error rate or misclassification rate is 19.1 percent.

3.4.3. Neural Network (NN)

The other data mining technique which will be employed in this study for the class prediction regarding the loan defaulter is Artificial Neural Network. The inspiration for the development of the Artificial Neural Network (ANN) is obtained from the biological concept of Neural Network. These networks are usually comprised upon input layer, a hidden layer and output layers with connections among each of them by the nodes or neurons. The application of neural network requires normalized data for a better performing model. After the normalization of the data, neural network technique was employed for the proposed model development. The employed neural network was comprised of 11 neurons in the input layer as the independent variables. While the first hidden layer has

The HPSPLIT Procedure

10-Fold Cross Validation Assessment of Model								
	Average Square Error				Misclassification Rate			
N Leaves	Min	Avg	Standard Error	Max	Min	Avg	Standard Error	Max
12	0.1274	0.1666	0.0288	0.2129	0.1333	0.2013	0.0422	0.2769

The HPSPLIT Procedure

Confusion Matrices				
		Predicted		Error Rate
	Actual	0	1	
Model Based	0	105	87	0.4531
	1	17	405	0.0403
Cross Validation	0	93	99	0.5156
	1	25	397	0.0592

Fig. 14. Confusion matrix of decision tree.

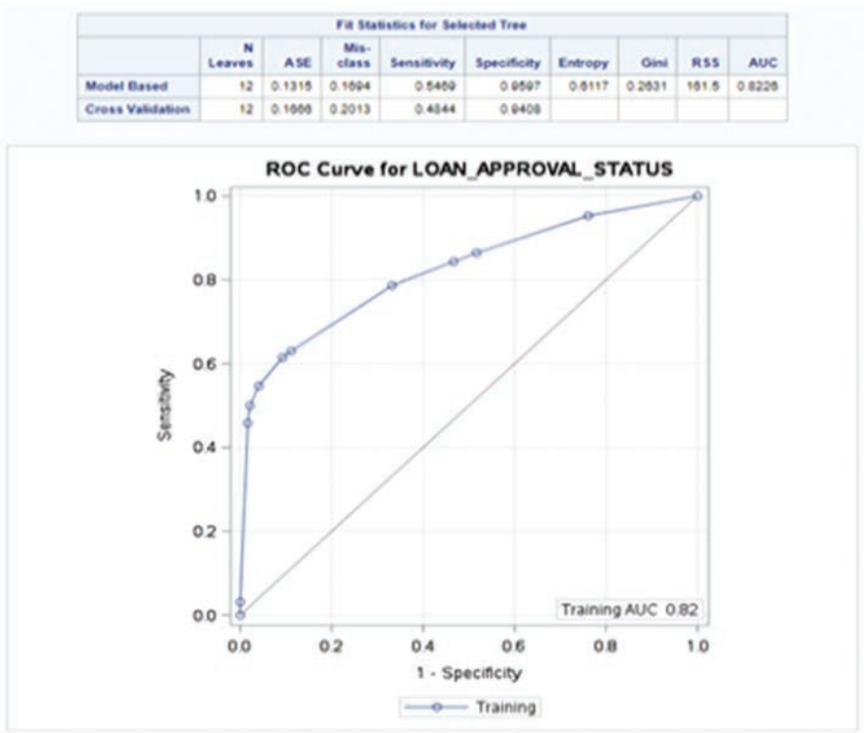


Fig. 15. ROC related to decision tree.

9 neurons and the second hidden layer was comprised upon 5 neurons. Furthermore, the last output layer has 2 neurons as the classifiers of the proposed model. Moreover, the architecture of the model was based on "Multi-Layer Perceptron" as MLP and 80 percent of the dataset was used for the training and 20 percent for the validation purpose. The implantation of the neural network was completed by employing "PROC HPNEURAL" function which can be observed in Figure 20 along with the structure of the developed model.

Furthermore, confusion matrix related to the validation model can be observed in Figure 21. The accuracy achieved by the neural network model was 83.07 percent with a misclassification rate of 16.93 percent.

Fig. 16. Structure of the logistic regression model.

3.5. Access

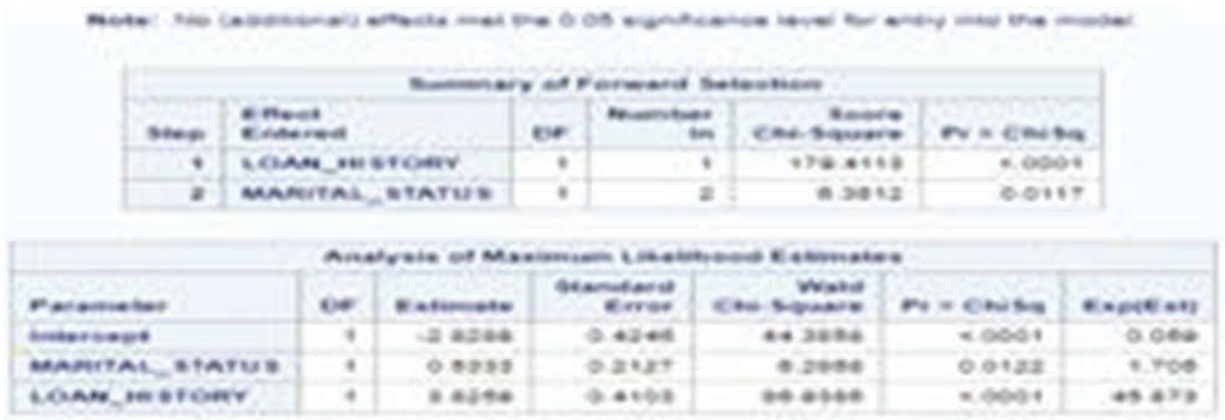
The final process of the SEMMA methodology after the model development phase is the model accessing process. This task is very crucial as every employed technique for the data mining or class prediction has to be evaluated and the reliability of the model has to access.

Once the model implementation from all the proposed techniques and validation has been satisfied. Testing data was incorporated in each model and the developed model was employed for the prediction of the loan approval by employing the trained model with all of the three techniques which are Decision Tree (DT), Logistic Regression (LR) and Neural Network (NN).

The trained decision tree model was employed on the testing dataset where the cut off probability was fixed at 0.5 to classify between the two classes of approval and disapproval of the loan applications. Initially, both training and testing datasets were appended for the logistic model assessment and then cut off probability was tuned at the value of 0.5 along with the forward selection technique for linear predictors. The assessment of the final model based on neural network was also conducted on the testing dataset. Although, the testing dataset was also normalized as the training was performed on the normalized data.

3.5.1. Assessment Parameters

In this study, confusion matrix will be applied for the model performance evaluation purpose. As confusing



Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	LOAN_HISTORY	1	1	179.6113	<.0001
2	MARITAL_STATUS	1	2	6.3812	0.0115

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept	1	-2.8266	0.4246	44.3856	<.0001	0.059
MARITAL_STATUS	1	0.5233	0.2127	6.2956	0.0122	1.708
LOAN_HISTORY	1	3.8256	0.4103	86.9368	<.0001	45.879

Fig. 17. Significant attributes of logistic regression model.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	59.7	Somers' D	0.488
Percent Discordant	10.8	Gamma	0.692
Percent Tied	29.5	Tau-a	0.210
Pairs	81024	c	0.744

Fig. 18. Association parameters of logistic regression.

Classification Table

	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.500	415	82	110	7	80.9	98.3	42.7	21.0	7.9

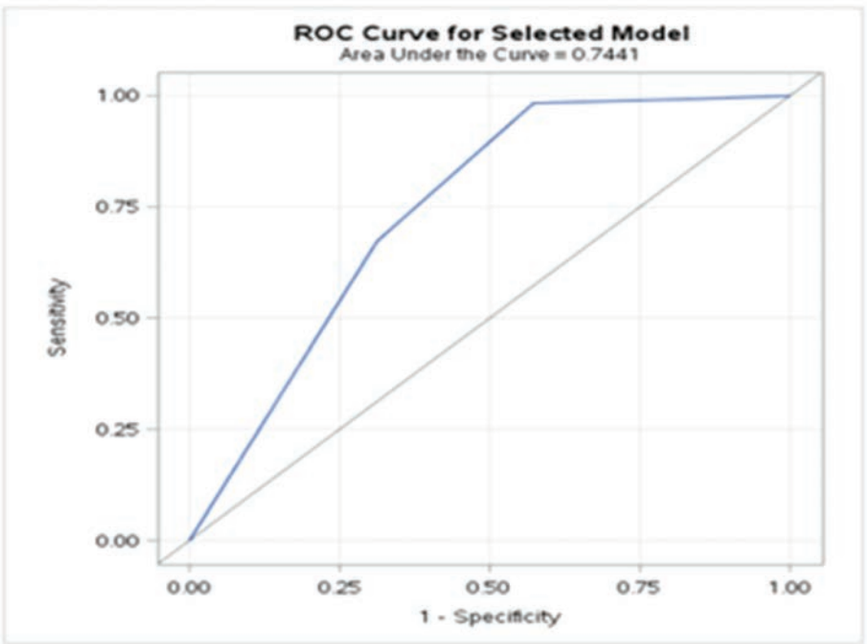


Fig. 19. ROC related to logistic regression.

Model Information	
Data Source	DAP.SME_TRAIN4
Architecture	MLP
Number of Input Variables	11
Number of Hidden Layers	2
Number of Hidden Neurons	14
Number of Target Variables	1
Number of Weights	184
Optimization Technique	Limited Memory BFGS

Number of Observations Read	614
Number of Observations Used	614
Number Used for Training	484
Number Used for Validation	130

Fig. 20. Structure of the neural network model.

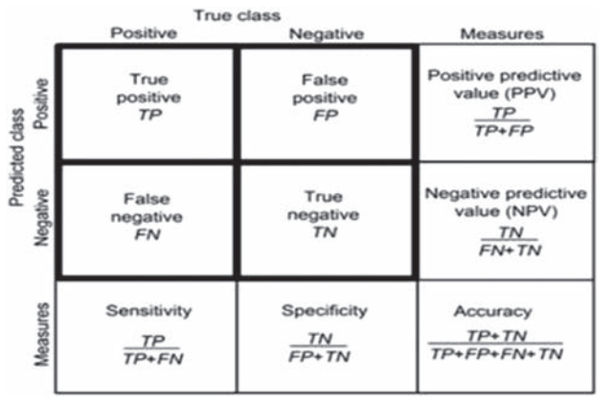


Fig. 22. Confusion matrix.

Table II. Parameters for evaluation.

	Parameter			
Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	Error rate (%)
Decision tree	79.8	78.8	80.04	21.2
Logistic regression	80.9	98.3	42.7	19.1
Neural network	83.07	83.4	80.95	16.93

matrix is a reliable method to summarize the performance of the classifier. Confusion matrix presents the number of predictions with respect to their correctness as the name suggested it shows that how confuse the classifier was during the prediction process. Tabular form is used in confusion matrix (see Fig. 22) for the purpose of the description of the classifier. From this confusion matrix, accuracy, error rate, precision and other significant rates [25].

It can be observed from Table II that in terms of loan defaulter predictive "Accuracy" parameter, neural network model performed best in contrast to the logistic regression and decision tree.

Although, the evaluation cannot be only depending upon the "Accuracy" as with respect to the proposed model which is regarding the loan defaulter prediction "Sensitivity" and "Specificity" parameters are most crucial. While the "Sensitivity" parameter represents that how well a model can predict "True Positive" values with respect to the "False Negative" ones. As in the financial world "True Positive" values are not considered as meaningful as "True Negative" values. So, if consider these significances with respect to the proposed model, Logistic regression have performed best in comparison to the decision tree and neural network although it does not have the highest "Accuracy."

Train: Number of Observations	Valid: Number of Observations	L1 Norm of Weights	Train: Average Error Function	Valid: Average Error Function	Train: Average Absolute Error	Valid: Average Absolute Error	Train: Maximum Absolute Error	Valid: Maximum Absolute Error
484	130	24.381508	0.478039	0.474247	0.305727	0.293972	0.959828	0.959260

Train: Number of Wrong Classifications	Valid: Number of Wrong Classifications	Train: Misclassification Rate	Valid: Misclassification Rate
96	22	0.1983	0.1692

Misclassification Table for LOAN_APPROVAL_STATUS		
Class:	1	0
1	91	4
0	18	17

Fig. 21. Fit statistics for neural network model.

4. CONCLUSIONS

This study has proposed a comprehensive research and model development for the prediction of the default loans. As the issue related to the high ratio of bad loans is very much critical in the financial sector especially in micro-financing banks of various under develop and developed countries. Although, loan lending has been proven very substantial in the stability of any country's economy in this century such a huge amount of loan defaults is also very critical.

To cope up with this problem a comprehensive amount of literature was reviewed to study the significant factors that lead to such problems. Moreover, these reviewed studies were critically focused towards the employed techniques and methods of data mining for the prediction and classification of the loan defaults. These data mining techniques include Neural Networks (NN), Support Vector Machine (SVM), Linear Regression (LR), Random Forest (RF), Decision Tree (DT), Logistic Regression (LG), Fuzzy Logic (FUZZY), Genetic Programming (GENETIC), Discriminant Analysis (DA), Bayesian Networks (BN), Hybrid Methods (HYBRID) and Ensemble Methods (COMBINED) which have been reviewed as well.

Furthermore, this study has also discussed the employment of these data mining techniques which were carried out by the implementation of a few significant methodologies. These methodologies include KDD, CRISP-DM and SEMMA. A critical performance comparison of these methodologies was conducted and SEMMA technique was selected for the model development phase. In the model development phase, all the five-key steps of SEMMA was elaborated before the deployment of the model.

While in the experimentation phase, three different data mining techniques were employed for the proposed model development and their performances were evaluated on various parameters. Based on these parameters, the best method was selected, explained and suggested because of its significant characteristics regarding the prediction of the loan defaults in the financial sector.

References

1. Ducai, M.T., **2012**. The bank loans importance, information asymmetry and the impact of financial and economic crisis on corporate financing. *Revista Tinerilor Economişti (The Young Economists Journal)*, (18), pp.29–34.
2. Wyman, O., **2015**. The role of financial services in society statement in support of macroprudential policies, Available at: http://www3.weforum.org/docs/WEF_The_Role_of_Financial_Services_in_Society_report_2015.pdf [Accessed: 12 June 2018].
3. Dunn, I., **2017**. Common Problems and Bottlenecks of Loan Portfolio Analysis | Visible Equity. Available at: https://www.visibleequity.com/common-problems-and-bottlenecks-of-loan-portfolio-analysis [Accessed: 12 June 2018].
4. Leung, A., **2017**. Here's a Look at the World's Worst Bad-Loan Ratios: Map-Bloomberg, Bloomberg. Available at: https://www.bloomberg.com/news/articles/2017-11-21/here-s-a-look-at-the-world-s-worst-bad-loan-ratios-map [Accessed: 12 June 2018].
5. Fay, B., **2017**. What is a Credit Score and How is it Calculated? Available at: https://www.debt.org/credit/report/scoring-models [Accessed: 12 June 2018].
6. Pritchard, J., **2018**. How Credit Scores Work. Available at: https://www.thebalance.com/how-credit-scores-work-315541 [Accessed: 5 August 2018].
7. Chye, K.H., Chin, T.W. and Peng, G.C., **2004**. Credit scoring using data mining techniques. *Singapore Management Review, 26*(2), p.25.
8. Abdou, H.A. and Pointon, J., **2011**. Credit scoring, statistical techniques and evaluation criteria: A review of the literature. *Intelligent Systems in Accounting, Finance and Management, 88*, pp.59–88.
9. Brown, I. and Mues, C., **2012**. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications, 39*(3), pp.3446–3453.
10. Kaggle.com, **2018**. Kaggle Data Repository [Online], Available: https://www.kaggle.com/ninzaami/loan-predication [Accessed: 24 October 2018].
11. Arun, K., Ishan, G. and Sanmeet, K., **2016**. Loan approval prediction based on machine learning approach. *IOSR Journal of Computer Engineering (IOSR-JCE), 5*, pp.18–21.
12. Louzada, F., Ara, A. and Fernandes, G.B., **2016**. Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science, 21*(2), pp.117–134.
13. Xu, J.J., Lu, Y. and Chau, M., **2015**. P2P Lending Fraud Detection: A Big Data Approach. *Pacific-Asia Workshop on Intelligence and Security Informatics*, Springer, Cham. pp.71–81.
14. Gupta, P., **2017**. Decision Trees in Machine Learning—Towards Data Science. [Online], Available: https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052 [Accessed: 17 June 2018].
15. Ma, L., Zhao, X., Zhou, Z. and Liu, Y., **2018**. A new aspect on P2P online lending default prediction using meta-level phone usage data in China. *Decision Support Systems, 111*, pp.60–71.
16. Hamid, A.J. and Ahmed, T.M., **2016**. Developing prediction model of loan risk in banks using data mining. *Machine Learning and Applications: An International Journal (MLAIJ), 3*, pp.1–9.
17. Gislason, P.O., Benediktsson, J.A. and Sveinsson, J.R., **2006**. Random forests for land cover classification. *Pattern Recognition Letters, 27*(4), pp.294–300.
18. Malekipirbazari, M. and Aksakalli, V., **2015**. Risk assessment in social lending via random forests. *Expert Systems with Applications, 42*(10), pp.4621–4631.
19. Gahlaut, A. and Singh, P.K., **2017**. Prediction Analysis of Risky Credit Using Data Mining Classification Models. *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, IEEE. pp.1–7.
20. Alomari, Z. and Fingerman, D., **2017**. Loan default prediction and identification of interesting relations between attributes of peer-to-peer loan applications. *New Zealand Journal of Computer-Human Interaction (ZJCHI), 2*.
21. Sarma, K.S., **2013**. Predictive Modeling with SAS Enterprise Miner: Practical Solutions for Business Applications. SAS Institute.
22. Nalić, J. and Švraka, A., **2018**. Using Data Mining Approaches to Build Credit Scoring Model: Case Study—Implementation of Credit Scoring Model in Microfinance Institution. *2018 17th International Symposium INFOTEH-JAHORINA (INFOTEH)*, IEEE. pp.1–5.
23. Luo, C., Wu, D. and Wu, D., **2017**. A deep learning approach for credit scoring using credit default swaps. *Engineering Applications of Artificial Intelligence, 65*, pp.465–470.
24. Ivan, M., **2018**. Classification using Random forest in R | en.proft.me. [Online], Available: https://en.proft.me/2017/01/24/classification-using-random-forest-r [Accessed: 20 July 2018].
25. Hesed.com, **2018**. Performance Evaluation. [Online], Available: http://hesed.info/blog/specificity-calculator.abp [Accessed: 20 July 2018].

Received: 24 October 2018. Accepted: 19 February 2019.