Prediction of Loan Scoring Strategies Using Deep Learning Algorithm for Banking System

Ashwani Kumar, Raman Dugyala, and Pronaya Bhattacharya

Abstract

There has been a tremendous growth in banking and finance sectors. With this growth, the ease to access of sanction loan has increased because many people are applying for loans. The problem here is that bank has only limited number of resources and capital, which the bank can distribute among the customers. The whole task of categorizing to whom the bank should sanction loan and to whom it should not has become a difficult task for the bankers. Generally, bank undergoes a rigorous procedure for verifying the customer to sanction loan. This procedure may take a week's time or two. The drawback here is that the customer needs to wait for two whole weeks to know whether he/she is deserving or not. In this paper, we have reduced the risking factor of banks behind finding the appropriate person for loan approval by the bank. We even reduce the time of loan approval analysis. We first use data mining techniques to analyze previous records to which the bank has already sanctioned loan based on the analysis made out of these records we train the deep learning model. The new data is treated as testing data, and the output of the customer is calculated accordingly.

Keywords

Deep learning algorithm • Data mining • Training & testing • Loan approval analysis

1 Introduction

Nowadays, distributing loans has been the integral business sector of almost every bank. The important aspect of assets used in banking directly comes from the profit earned from the loans. The best step of any banking system is to identify the worthy stakeholders from which they can get maximum profit from the investment in the assets. The field of banking is affecting the lives of the loan holder by its services. Financial companies grant loan after a rigorous process of verification and validation, but still there is no surety that the loan being granted to the deserving candidates out of all the applicants. Through in our proposed approach, we can predict whether that particular applicant is provided loan sanction service or not. The entire process of validation of features is calculated automatically by deep learning algorithms.

Our paper prediction of loan scoring strategies using deep learning classifier for banking system provides a solution for the bank employees as well as for the applicant who is seeking for a loan. In this paper, we provide a simple and fast way to the bank employees to choose the deserving applicant from the number of applications. The proposed approach can calculate the values of every parameter taking part in loan scoring strategies on new test data same parameter are processed with respect to their related values. We also used time constraint which can be set for all loan application to check whether the loan can be approved or not for a person. In the current situation, the success and failure of banking system directly depend the analysis of credit risk. If the bank could not collect the amount back from the loan person properly, this will lead a loss to bank. Hence, calculating credit risk plays a vital role to be managed by the bank (Sudhamathy & Jothi Venkateswaran, 2016). Loan scoring strategies are two types: First is application scoring and the second one is behavioral scoring. In application scoring, classifiers are identified from the loan applications to calculate score. The classifiers are 'good', 'bad', 'risk',

A. Kumar (✉) · R. Dugyala
Sreyas Institute of Engineering and Technology, Hyderabad, India
e-mail: ashwani.kumarcse@gmail.com

P. Bhattacharya
Institute of Technology, Nirma University, Ahmedabad, India

'high risk' group. In the behavioral scoring, the classifier depends on the customer payment history and personal information (Babu & Satish, 2013; Laha, 2007). The important part of the banking industry is to measure and minimize the risk associated with a financial loss. For calculating the risk, our model uses risk evaluation model (Arutjothi & Senthamarai, 2016, 2017). The use of classification and prediction is the important key points for the proposed approach. In the current scenario, deep learning-based methods provide a very good accuracy for prediction models. In our proposed work, we will be using the decision tree techniques of deep learning to build this prediction model to predict loan scoring analysis because decision tree gives very good accuracy in the prediction.

The prime objective of this paper is we have to use data mining techniques to analyze previous records to which the bank has already sanctioned loan. Based on this analysis, we trained our deep learning-based model to predict the loan for a decision. The main objective of this paper is to predict whether the loan can be sanctioned to a person or not. To support the proposed approach, we are using data collection, deep learning models, training of the proposed model, logistic regression and testing. We have also compared different machine learning classification models based on the data collected and selected the best model, which gives a good accuracy. In this paper, the proposed deep learning algorithm is used to check whether the person can avail the facility of loan or not by calculating the data with the help of logistic regression classifiers which gives the accurate result for the prediction. Our proposed approach gives benefit to both parties for customer; it reduces the time period of loan approval and for bank employees it reduced the risking factor of bank in behind finding the appropriate person for loan approval by the bank.

This research paper is organized as follows: Sect. 1 shows the introduction of the loan prediction and deep learning algorithm. The Sect. 2 provides the literature review of deep learning and prediction models. In the Sect. 3, we have shown the proposed work. Section 4 gives result analysis and scope of the research in the fields of loan prediction. Finally, the Sect. 5 concludes the research article.

2 Literature Survey

In the literature review, we have gone through various research papers for prediction of loan. We find many data mining algorithms for the prediction of loan using different tools. We also read the research papers in which the minimum required parameter explained for the prediction. We have identified risk assessment and entropy the important issues in the financial institutions such as banking system. Prediction of loan scoring strategies is widely analyzed using

classification methods. It uses feature selection technique to remove the irregularity of the attributes. Abddmoula applied K-NN classifier on the Tunisian commercial loan dataset which gives about 88.63% of classification rate. Bach (Abdelmoula, 2015) in his paper highlights on loan decision-making systems with several feature selection techniques and classifiers in which evaluator-based system provides highest accuracy than other feature selection techniques. Arutjothi proposed a new credit scoring model, which uses the hybrid feature selection method which has a strong mathematical basis, but also has higher accuracy and effectiveness. Nikhil Madane et al. proposed a paper loan prediction using decision tree they make use of decision tree induction algorithm for implementing a model and review credit scoring of mortgage loans for the applicants. This credit score helps in sanctioning of the loan; hence, its assessment is mandatory. The model is used to predict a safe for loan sanctioning using Kaggle dataset. Om Prakash Yadav et al. proposed a paper on loan prediction in which they have tried to evaluate the credit risks and to identify the loan repayment prediction using decision tree algorithm (Bach et al., 1997; Kumar & Goel, 2020). Arun et al. (2016) focus to reduce the efforts of bank employees by generating a model by various machine learning algorithms and explained which of the methods can be accurate. To sum up in one line, all research articles explained in literature review use different algorithms such as min-max normalization, KNN algorithm for the prediction of loan which gives 75.08% accuracy result.

Logistic regression is used to predict the probability of an outcome that can only have two values (Madane1 & Nanda, 2019). The prediction is based on the use of one or several predictors. It produces a logistic curve, which is limited to values between 0 and 1. Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. For data cleaning process, various methods are used; one of the important method is Bayesian technique. In this method, missing values can be filled by calculating nearest neighbor values and can be identified by regression. Regression is a technique which uses Bayesian classification along with decision tree. This method is the first choice of the researcher for predicting the missing values. By making a good decision, it results in finding out more accurate data entry for missing values (Nisbet et al., 2009; North, 2012; Pujari, 2001). To convert large volume of dataset into smaller volume, data reduction methods are used which preserve data integrity (Han et al., 2012). The dataset may contain various repeated attributes so there is a need for removing the redundant attributes. The selection of necessary attribute is done by a mechanism known as feature selection which is used to decrease the dataset by discarding the irrelevant attributes (Gupta, 2014; Raudenbush & Anthony, 2002; Whitney 1971; Witten et al. 2016). In our
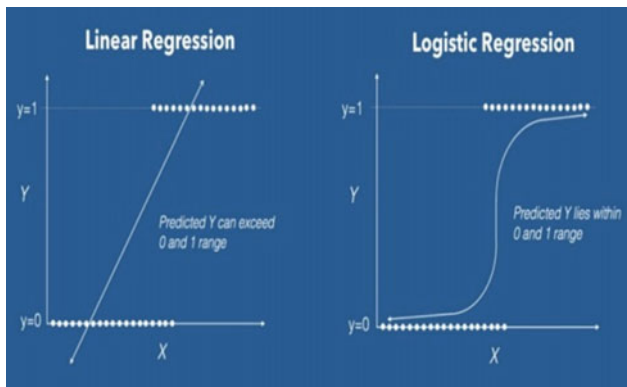
Fig. 1 Linear regression versus logistic regression

country India, the number of loan application drastically increased in recent years. The major problem experienced by the bank employees is they are not able to take decision about loan approval or analysis whether the customer can pay back the amount or not. Currently, all commercial banks are trying to find out effective way to motivate customers for applying their loans. The other problem is with the sanctioned loan, there are some customers who did not shows a positive response once their loan application gets approved. Our proposed approach gives a solution to prevent these situations; banks must find some models to predict customer's behaviors. Machine learning-based techniques give a good performance for this purpose. Figure 1 shows the comparison difference between linear regression and logistic regression.

There is a need for increasing the accuracy in prediction of loan approval in banking sector. Therefore, this paper presents a deep learning-based algorithm by using logistic regression along with minimizing the risk factor for the prediction of loan in commercial banks. We cannot say which model is best because model has its own specification and accuracy depends on data provided to it. Kumar et al. (Kumar & Srivastava, 2020; Kumar et al., 2019) proposed an object detection method for blind people to locate objects from a scene. They have used machine learning-based methods along with single SSMD detector algorithm to develop the model (Raman & Krishna, 2013; Dugyala et al. 2016).

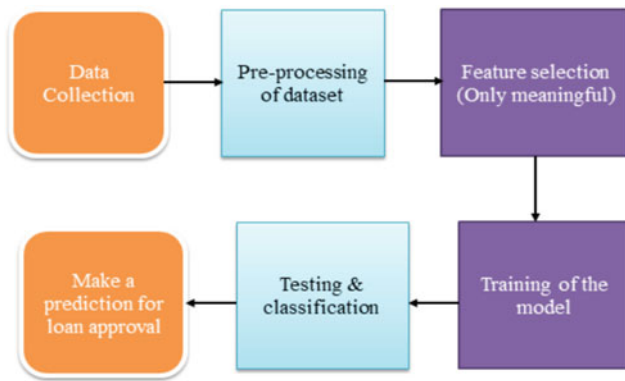# 3 Loan Scoring Strategies Using Deep Learning

The model used in our research work focuses on predicting the credibility of user for loan repayment by analyzing their behavior. Customer behavior is the input provided into the model, and the output is a decision whether to sanction or reject loan using data analytics tools. The process of predict loan approval required to train the data using deep learning algorithms and then compare past customer data with trained data for making a decision. The collected datasets may contain missing values, corrupted data and anomalies in the data, which needs to be properly managed or discarded. The collected dataset uses many attributes that define the behavior of the customers. In this section, we have provided the prediction of loan scoring strategies using deep learning algorithm for banking system for that we have used important parameter such as risk analysis, entropy and logistic regression. Figure 2 shows the system model used for prediction.

In this approach, we are using data collection, deep learning models, training of the model, logistic regression, dataset, validation and testing. Our model will help the bank employees to predict the trustworthy persons who have applied for a loan, thus increasing the chances of retaining their loans in time by using decision tree. The process includes first the bank manager will verify the customer's eligibility for loan using proposed model while giving some basic information of the customer. Second the customer can also check loan eligibility by providing required information, and in response, he will receive email whether he is eligible for loan or not. Hence, there is no need for the customer to visit bank many times leads in time saving for the customers. The entire process of loan prediction can be understood form the system model shown in Fig. 3.

Figure 3 shows a model used for prediction of loan scoring strategies using deep learning algorithm. First, data collection is performed for training and testing purpose for loan prediction. Second, after data collection the trained model is prepared using the k-means algorithm. Lastly, we have provided the test datasets for the model to make a
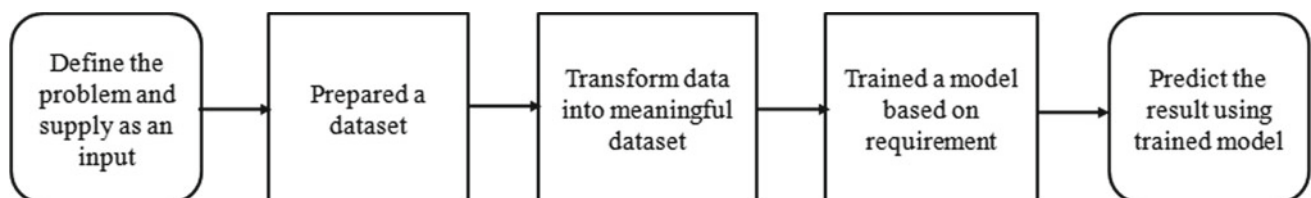
Fig. 2 System model used prediction

Fig. 3 Model used for prediction of loan scoring strategies using deep learning algorithm

prediction. The scheme is used to identify the defaulters who did not pay loan on time. K-means algorithm helps in the process of classification of dataset.

3.1 Data Collection

Data collection refers to the previous data records of the customers to whom the loan was sanctioned collected.

3.2 Training of Model

The collected data given to deep learning model as input is termed as training data. In this, we analyzed various fields of collected data and find out the attributes used in the records. We first trace out the missing data and fill them with appropriate values.

3.3 Testing

We supply the new records to deep learning model and fetch output for the same. The output obtained would be based on the conclusions made out of previous data.

3.4 Dataset

The dataset includes different types of training dataset, validation dataset and test dataset. In the training dataset, sample of data is used to fit the model. This is the actual dataset that we use to train the model. The validation dataset is sample of data used to provide an unbiased evaluation of a model. The test dataset is the gold standard used to evaluate the model.

3.5 Validation

In validation dataset, we consider various parameters such as loan status as Yes or No. For testing, we consider some sample data and test the data to check the desired output.

3.6 K-Means Algorithm

Applying of k-means yields k which supports the original n data points. These are more similar to each other that belong to the other clusters. This technique is used for dynamic clustering.

$$J(v) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\| x_i - v_j \| \right)^2 \qquad (1)$$

The following are steps used in the process of loan prediction.

|	
Step 1	We have to load dataset into the classification model
Step 2	Perform the pre-processing on the dataset. The data sets may contain missing values, corrupted data and anomalies in the data
Step 3	Classify the customer into different cluster classes based on customer behavior
Step 4	Feature selection performed on the dataset while considering certain important attributes
Step 5	Build a prediction model using deep learning algorithm
Step 6	Make a prediction for loan approval.

In loan scoring strategies using deep learning, a data mining technique is used for identifying the classifier. This classifier can make a good decision from the input set to correctly predict an accurate class based on trained dataset and should be able to learn complex patterns. The proposed scheme consists of data collections, pre-processing, feature selection and prediction phases. Deep learning-based algorithms are one of the mostly used methods for loan prediction (e.g., worthy, not worthy, cannot say and risky) (Kumar et al. 2020; Yadav & Soni, 2019). We had developed deep learning-based model which uses K-means algorithm for decision making for making a prediction. The main aim of the proposed model is to predict loan approval with high accuracy.

4 Result Analysis

In this section, we have discussed the result of our proposed approach. We can predict the status of a loan application for the customer applied for the loan by providing all the
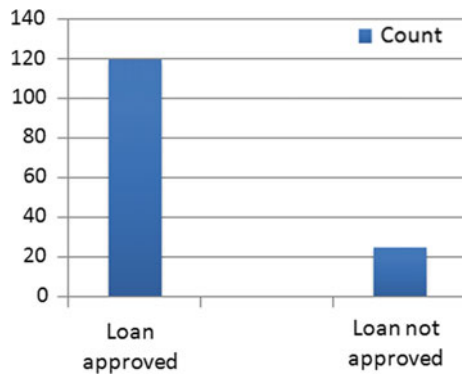
Fig. 4 Status of loan to approve or not using training dataset
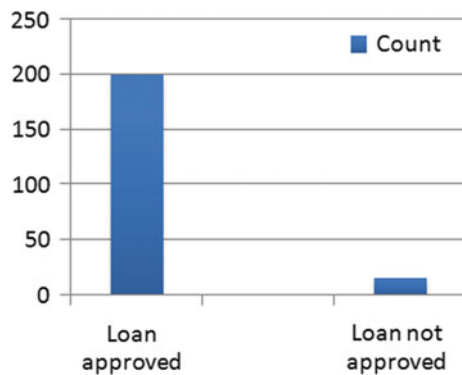
Fig. 5 Status of loan to approve or not using actual dataset

attributes into our model. We also have test cases in which we predict the number of customer deserving for loan sanction by supplying the testing data to our model. Since we have used the logistic regression algorithm, the output of the data would be binary, i.e., yes or no or 1 or 0. The testing data is stored in a variable called prediction variable.

Figures 4 and 5 demonstrate the performance of the proposed model using test dataset and actual dataset. The model uses k-means algorithm to take decision and logistic regression to detect accurate result. The model also tells us whether to approve or reject a loan application applied by a customer. This model stores a table of trustworthy and defaulter customer from the previous datasets. The bank employees can use this model to reduce the risk of

investment failure by providing loan services to the defaulters and even reduce the time period of loan approval analysis. Our model gives correctly 81.3% performance while applying on test dataset.

Figure 6 represents histogram graphs for the various attributes for the loan approval. These histogram graphs contain number of people on the y-axis and loan approval attribute on the x-axis. These graphs give us a rough idea that how the customer details are directly or indirectly related to loan sanction. These graphs are plotted using Matplotlib library. Table 1 represents the description of each features used in the dataset. These feature selection is very crucial and used to judge the behavior of the customer applied for loan approval. The proposed model can predict whether to approve or reject the loan application. Heap map is used which describes the attributes of customer for loan prediction, and they are correlated with each other. The graph describes attributes like credit history, loan amount term, loan amount, etc.

5 Conclusion

In this paper, we have provided a solution for bank employees to predict loan approval for a particular customer using the trained model. The prime aim of prediction of loan scoring strategies is to classify and analyze the nature of the loan applicants. The graphs generated during data visualization phase give more clarity about the prediction. These graphs also provide the information about short-term loan identified, as most preferred by majority of the loan applicants. Prediction of loan scoring strategies helps banking sector for predicting the future of loan and its status. This enables bank employees to take action in the initial days for loan approval. The proposed approach would be very useful to the banking system for better targeting and acquiring new customers. The said model gives vital information to avoid a big financial loss for the banking institutions. The use of deep learning-based technique in the proposal model gives a very good accuracy. In future, efficient machine learning algorithms can be used to further increase the accuracy and retrieved precise results of the model.
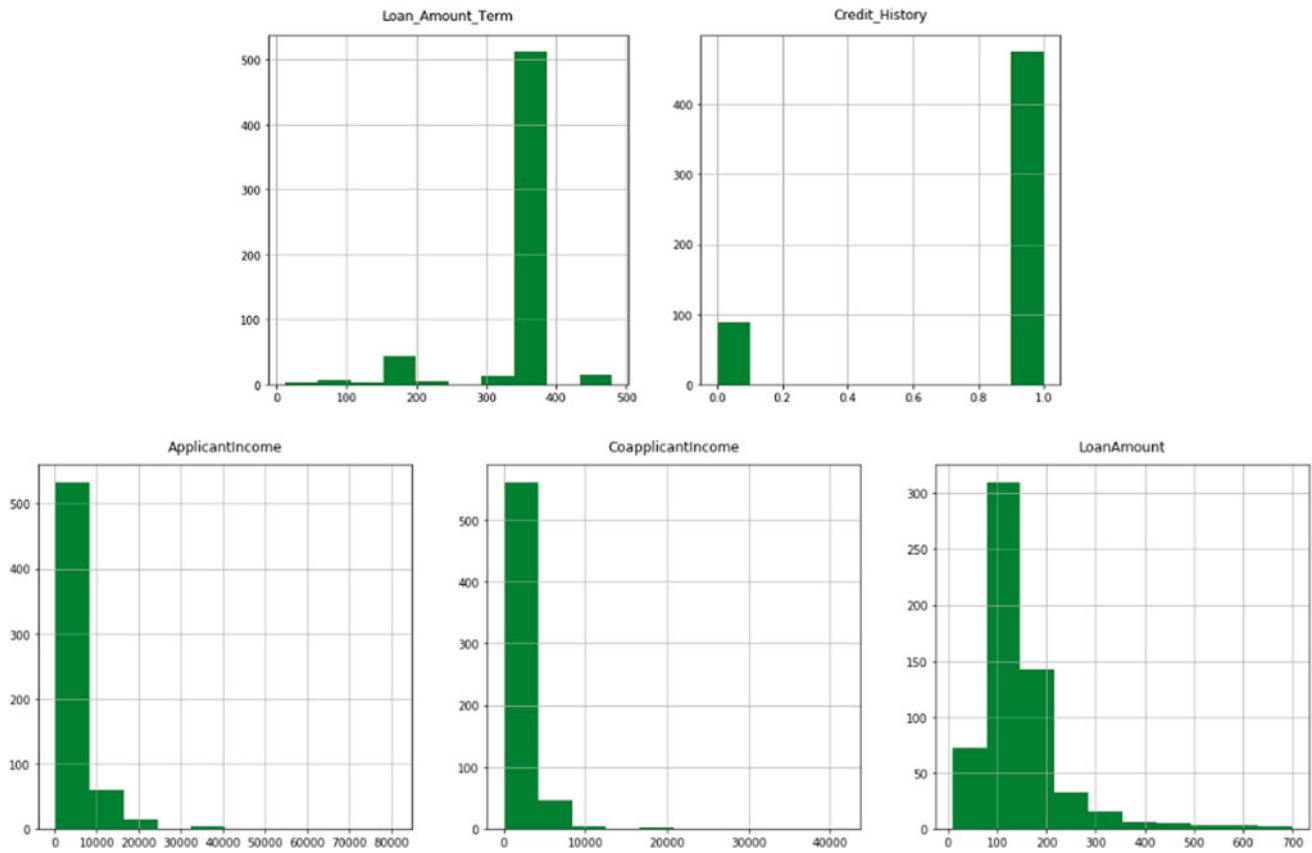
Fig. 6 Histogram graphs of various attributes of loan approval

Table 1 Description n of each features used in the dataset

S.R.	Feature selection based on customer behavior	Description of features
1	loan_policy	1 If the customer meets the loan underwriting criteria, and 0 otherwise
2	type_of_purpose	This refers to the purpose of the loan
3	int_rate	The interest rate of the loan
4	days_	The no. of days to loan sanctioned
5	inq_	The loan holder inquiries
6	no_of_installment	The monthly installments owed by loan holder if the loan is sanctioned
7	loan_fully_paid_or_not	This indicates whether the loan was fully paid *r* or not

References

Abdelmoula, A. K. (2015). Bank credit risk analysis with k-nearest-neighbor classifier: Case of Tunisian banks *Accounting and Management Information Systems*, *14*(1).

Arun, K., Ishan, G., & Sanmeet, K. (2016). Loan approval prediction based on machine learning approach. In *National Conference on Recent Trends in Computer Science and Information Technology (NCRTCSIT-2016)*.

Arutjothi, G., & Dr. Senthamarai, C. (2017). Comparison of feature selection methods for credit risk assessment. *International Journal of Computer Science*, *5*(I), No 5.

Arutjothi, G., & Senthamarai, C. (2016). Effective analysis of financial data using knowledge discovery database. *International Journal of Computational Intelligence and Informatics, 6*(2).

Babu R., & Rama Satish, A. (2013). Improved of K-nearest neighbor techniques in credit scoring. *International Journal For Development of Computer Science & Technology* I.

Bach, M. P., et al. (1997). Selection of variables for credit risk data mining models preliminary research. *MIPRO 2017—40th Jubilee International Convention*. Aldonado, M., Chang, C.-C. K., Gravano, L., & Paepcke, A. (1997). The stanford digital library metadata architecture. *International Journal Digital Library*, *1*, 108–121

Dugyala, R., Bezawada, B., Thatiparthi, R. K. V., & Sathyanarayan S. (2016). Static program behavior tracing for program similarity quantification. In *Proceedings of the First International Conference on Computational Intelligence and Informatics Vol. 507 of the series Advances in Intelligent Systems and Computing* (pp 321–330).

Gupta, G. K. (2014). *Introduction to data mining with case studies*. Delhi, India: PHI Learning Pvt. Ltd.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining concepts and techniques*. USA: The Morgan Kaufmann Series in Data Management Systems, Elsevier.

Kumar, A., Reddy, S. S. S. S., & Kulkarni, V. (2019). An object detection technique for blind people in real-time using deep neural network. In *2019 Fifth International Conference on Image Information Processing (ICIIP)*, Shimla, India (pp. 292–297). https://doi.org/10.1109/ICIIP47207.2019.8985965.

Kumar, A., & Srivastava, S. (2020). Object detection system based on convolution neural networks using single shot multi-box detector. *Procedia Computer Science, 171,* 2610–2617.

Kumar, A., Zhang, Z. J., & Lyu, H. (2020). Object detection in real time based on improved single shot multi-box detector algorithm. *EURASIP Journal on Wireless Communications and Networking, 2020*(1), 1–18.

Kumar, S., & Goel, A. K. (2020). Prediction of loan approval using machine learning technique. *International Journal of Advanced Science and Technology*, *29*(9s).

Laha, A. (2007). Building contextual classifiers by integrating fuzzy rule based classification technique and k-NN method for credit scoring. *Advanced Engineering Informatics, 21*(3), 281–291.

Madane1, N., & Nanda, S. (2019). Loan prediction analysis using decision tree. *Journal of The Gujrat Research Society, 21*(14s).

Nisbet, R., Elder, J., & Miner, G. (2009). *Handbook of statistical analysis and data mining applications*. Canada, UK: Academic Press Publications, Elsevier.

North M. (2012). *Data mining for the masses*. Global Text Project.

Pujari, A. K. (2001). *Data mining techniques*. UK: Universities Press, Sangam Books Ltd.

Raman, D., & Krishna, B. V. (2013). Ensuring security services for data storing and data sharing in cloud computing. *International Journal of Science and Research (IJSR), 2*(2). ISSN 2319-7064.

Raudenbush, S. W., & Anthony, S. B. (2002). *Hierarchical linear models: Applications and data analysis methods*. USA: Sage Publications.

Sudhamathy, G., & Jothi Venkateswaran, C. (2016). Analytics using R for predicting credit defaulters. In *IEEE International Conference on Advances in Computer Applications (ICACA)*. IEEE.

Whitney, A. W. (1971). A direct method of nonparametric measurement selection. *IEEE Transactions on Computers, 100*(9), 1100–1103.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques*. USA: Morgan Kaufmann Series in Data Management Systems, Elsevier.

Yadav, O. P., Soni, C., Kandakatla, S. K., & Sswanth, S. (2019). Loan prediction using decision tree. *International Journal of Information and computer Science, 6*(5).