Springer Open

Modern Approach for Loan Sanctioning in Banks Using Machine Learning

Golak Bihari Rath, Debasish Das, and BiswaRanjan Acharya

1 Introduction

Loan analysis is a process adopted by banks used to check the credibility of loan applicants who can pay back the sanction loan amount within regulations and loan amount term mentioned by the bank. Most banks use their common recommended procedure of credit scoring and background check techniques to analyze the loan application and to make decisions on loan approval. This is overall a risk-oriented and a time-consuming process. In some cases, people suffer through financial problems while some intentionally try to fraud. As a result, such delay and default in payment by the loan applicants can lead to loss of capital of the banks. Hence to overcome this, banks need to adopt a better procedure to find the trustworthy applicants for granting loan from the list of all applicants applied for the loan, who can pay can their loan amount in stipulated time [1].

In the modern-day age and advance of technology, we adopt a machine learning approach to reduce the risk factor and human errors in the loan sanction process and determine where an applicant is eligible for loan approval or not. Here, we examine various features such as applicant income, credit history, education from past records of loan applicants irrespective of their loan sanction, and the best features are determined and selected which have a direct impact on the outcome for loan approval [2].

G. B. Rath (✉) · D. Das
Department of Computer Science & Engineering, Centurion University of Technology & Management, Bhubaneswar, Odisha, India
e-mail: golakk.2009@gmail.com

D. Das
e-mail: debasish.das@cutm.ac.in

B. Acharya
School of Computer Engineering, KIIT Deemed to University, Bhubaneswar, Odisha, India
e-mail: acharya.biswa85@gmail.com

S. Patnaik et al. (eds.), *Advances in Machine Learning and Computational Intelligence*,
Algorithms for Intelligent Systems, https://doi.org/10.1007/978-981-15-5243-4_15

Various machine learning algorithms such as logistic regression, decision tree, and SVM have been tested, and their results have been compared. The performance of logistic regression was found more than other models, and hence, it was assumed that it could be used as a predictive model which could predict future payment behaviors of the loan applicants. Thus, the bank could adopt this model for loan sanction process whenever new applicants apply for a loan, and the loan can be processed instantly with minimum time and reduced risk.

2 Literature Review

Zurada J. (2002) found ensemble model performs better in comparison with other data mining techniques by exploring the application details of both paid and defaulters [1]. Turkson et al. (2016) analyzed the performance of 15 different classification methods with 23 different features of applicant data and found that linear regression was used to formulate as the final model [2]. Vaidya, A. (2017) found out taking more number of attributes will result in the model learning better using logistic regression techniques by statistically analyzing the distribution of features and prediction of the model [3]. The work proposed by Hamid and Ahmed (2016) depicts the data mining process for loan classification using Weka application. Algorithms such as j48, Bayes Net, and Naive Bayes were used for model creation. Results showed j48 had the highest accuracy and low mean absolute error so considered the best suited for prediction [4]. Sivasree and Rekha Sunny (2015) used the Weka explorer tool for data exploration and implementation of using decision tree induction algorithm to find relevant attributes in decision making. ASP.NET-MVC5 was used as the platform for implementing the model into the application for use [5]. The authors Arun et al. (2016) explained the application of six algorithms applying parameter setting using R open-source software [6]. Shin et al. (2005) presented the prediction of the bankruptcy prediction model using support vector machine and good classifier to attain correct prediction performance from smaller sets [7]. Salmon et al. (2015) measured the performance scores of different classifiers using the evaluation technique of the confusion matrix [8]. Panigrahi and Palkar (2018) used various feature selection techniques for various models to determine fraud claims and found random forest model has the best accuracy and precision, whereas decision tree has the best recall using various feature selection methods applied on the dataset [9]. Soni and Paul (2019) compared the results of the model in R and Weka and found Weka results were better for making an optimized random forest classifier [10].

Arutjothi and Senthamarai (2017) proposed a K-NN credit scoring system using where the error rate is recorded minimum when the iteration level is increased [11]. Priyanka and Baby (2013) suggested a Naive Bayesian algorithm for classifying a customer according to posterior probability using records of customers in banks [12]. Sudhamathy G. (2016) implemented R package for visualization using data mining techniques. For the prediction of labels, the tree model was used [13]. Eweoya et al.

(2019) found that incorrect predictions can be reduced with using stratified cross-validation on a decision tree model [14]. Jency et al. (2019) preferred using an exploratory data analysis approach for graphical representation and understanding of different features of the people. It was found that short-term loans were preferred more and people having a home as mortgage apply more for a loan [15]. Rawate and Tijare (2019) used a Naive Bayes approach along with a combination of various algorithms such as K-NN, binning for consistency in dataset and improvement in the accuracy rate [16]. Priya et al. (2018) concluded that loan proposals of people with good credit history with high income have a better chance of approval [17]. Torvekar and Game (2019) experimented with various classifiers on two different datasets, and classifiers performance is reduced when operating with a large number of features [18]. Singh and Chug (2017) suggested that the calculation of the error rate metric is needed to find the best algorithms whenever various algorithms have the same accuracy rate. The linear classifier was found best for determining a software defect using tenfold cross-validation techniques [19].

In our paper, we try to use a similar machine learning approach exploring 12 different features of a banking dataset which affect the loan approval directly in some manner. We then apply feature selection techniques on the existing dataset to manually select those features based on weighting scores which contribute most to predicting a category or outcome using certain techniques of feature selection. We use the scikit-learn library which provides the SelectKBest class using the chi-squared (chi^2) test to select a specific number of features [20]. We select seven relevant features from 11 different attributes of previous loan applicants. Finally, the selected best features are fed into classification models for the classifying the outcome.

3 Proposed Model

To achieve our objective, we train the system with two parameters. First parameter is the predictive features referred as independent variables, and the other referred as categorical class for those variables. This system is a predictive model to determine the class as yes for approve and no for disapproval of loan dataset of loan applicants. Data is collected from a banking dataset containing recent records of applicants whose loan application has been either approved or disapproved. Supervised machine learning techniques are performed on that historical data to create a proposed model which can forecast the identification of loan applicant's repayment. We used scikit-learn library which supports Python functions in this process to visualize the distribution of data features and creating a classification model [20]. As this is classification problem of supervised learning, so we use algorithms such as logistic regression, decision tree, and SVM which can be best effective to solve the classification problem by categorizing the class for loan applied by the applicant is either risky or safe.

We implement a two-step process: model learning and model prediction. In model learning, a model is developed by using different classifier functions for the construction of a classification based on given training data. In model prediction, the model is used to predict the response for given data. Initially, we train a dataset with a set of features as input or predictor variables, and with an outcome variable Loan Status. All models are constructed and trained for classification with the help of classification algorithms. The logistic regression model is finally used as the best model for classification after the evaluation results of the prediction.

3.1 Logistic Regression

It is a statistical-based algorithm that determines the probability of an event as a function of variables using classification function. The classification function calculates the statistics for a logistic response function. It shows the relationship between the dependent variable and independent variables [1].

$$P(y) = 1/1 + \mathrm{e}^{-z} \tag{1}$$

Here $P(y)$ is our result which is determined with the help of dependent variable y, where z represents the function of independent variables used in the dataset. The range of values that $P(y)$ predicts is from 0 to 1 which helps us to identify our category as no or yes as results [1].

3.2 Decision Tree

Decision tree uses the tree representation to produce a model with the most relevant predictor attributes. Attribute with the highest-ranking attribute of the dataset is placed as the root node, and other attributes are placed to the leaf node. At each node, a decision is made where leaf nodes give us the final result. The tree building is continued until we get an outcome in the internal nodes. The overall results are calculated in the form of decisions, and the final decisions constitute our category [5].

3.3 SVM

Support vector machine is a classifier which is represented by a line that splits the data between the two differently classified groups of data representing the training data into two-dimensional planes as data points. We plot the features of the dataset in either side of the plane forming two different categories. When the testing data lands on either side of the line, we can classify the data for approval as yes or no [7].

4 Experiments and Results

Data is collected in the form of banking credit dataset containing records of past loan applicants from the UCI machine repository. There is a definite set of inputs defined as independent variables and a corresponding output referred as dependent variable or class. As the class is binary in nature, so we find a solution to the binary classification problem by adopting various classification techniques of machine learning such as decision tree, logistic regression, and SVM with the features of the dataset, and the best performance is calculated among them to create a predictive model. We use Python with the help of reading the data, exploring with visualizing different features of an existing dataset, and implementation of the model for prediction.

We implement our methodology with the processes of data exploration, data preprocessing, feature selection, model training, and model evaluation. Initially, we extract the data from CSV dataset file to pandas data frames using pandas class with read_csv() function. After reading the records and having a general understanding about various roles of different variables in the dataset, we prepare the data for building a model and prediction of results using various functions supported by scikit-learn [20].

4.1 Data Exploration

Before making a prediction model, we try to have a basic idea about both categorical and numeric data of the current dataset containing a vast number of data. The data in tabular format is displayed in graphical format by exploring the variables for calculation of frequency. Frequency of categorical variables is represented in Fig. 1 as a bar chart.

Here, we observe that the maximum number of applicants applied for the loan are male, graduates, employed, and married. Visualization of numerical variables like loan amount and applicant income is displayed in Figs. 2 and 3.

We can see that there is a considerable amount of outliers in the numerical variables along with missing values in categorical values which can be resolved in the next phase of preprocessing.

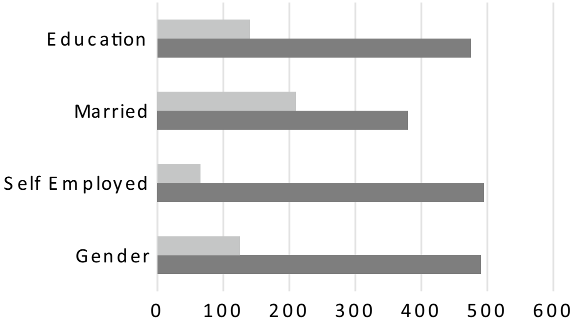

Fig. 1 Distribution of categorical values in chart

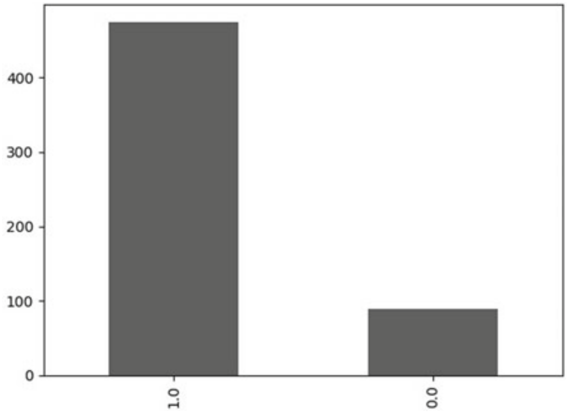

Fig. 2 Bar chart displaying credit history with 1 as yes and 0 for no

4.2 Data Preprocessing

The preparation and cleansing of the dataset are performed in this processing phase. It includes data cleansing which involves detecting and correcting (or removing) corrupt or inaccurate records from the current dataset. We check the missing data and fill them with mean or median values. Next, we perform numeric conversion of all categorical data types. The scikit-learn library supports only numerical variables during processing, hence we convert all categorical variables into numerical data types by encoding. We check that our data is now complete and numeric in nature which is suitable for other processing phases [20].

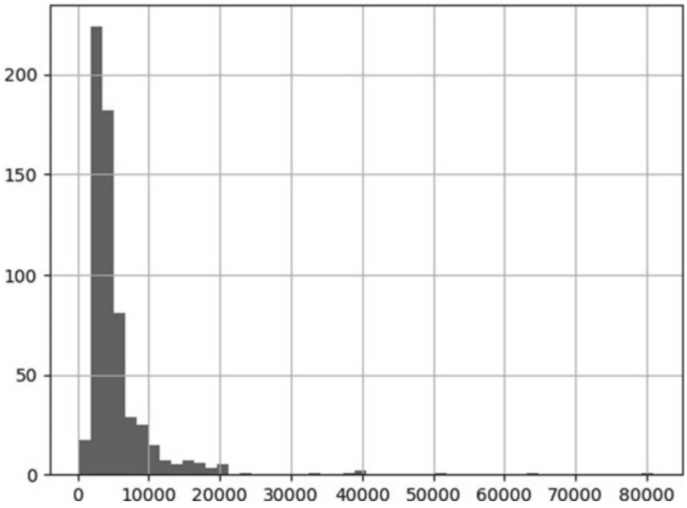

Fig. 3 Histograms displaying annual income of applicants

4.3 Feature Selection

When all features are not related to the outcome of a class, hence we select the relevant features with the help of technique such as feature selection. It helps in increasing the interpretability of the model and reducing the complexity and the training time of the model. Feature selection is performed by a class called SelectKBest provided with the scikit-learn library to identify the most important features which impact the loan approval results. This class uses a chi-squared (chi^2) test to determine the relative importance of independent features with the outcome variable [20].

In Table 1, we can see the output of the test and find seven dominant features scores of variables related to target variable 'Loan_Status'. So accordingly we choose some predictors in order of their weight to fit into the model.

Table 1 Feature score of variables

Feature	Score
Applicant income	5342.194844
Co-applicant income	4988.307182
Loan amount	110.4371
Credit history	19.617746
Married	2.132101
Education	1.793838
Dependents	0.806228

4.4 Training a Model

Once all predictive variables are selected, we need a test dataset for evaluating the performance. Hence, we split the current dataset into two sets, one for training and the other for testing. We randomly split the data with the help of scikit library supported train_test_split function. Thus, random sampling is performed using train_test_split function with three parameters such as predictors, outcome variable, and test size [20]. Our existing dataset contains 615 applicants. We split this dataset into the training and the testing datasets in certain proportions with the test size parameter. We split this dataset into the training and the testing datasets in certain proportions like 70% of the data for training the model and 30% of the data as testing dataset. We store predictors in x and the target variable in y separately in an array. We try with a different set of instances and predictors using different classification algorithms. First, we create an object of these classifiers to build the model by using its classification function for the algorithm and then fit the model with our parameters. Now the model starts to learn from the training data and is ready to predict for the testing dataset. Using the predict function, we predict and compare the results.

4.5 Model Performance Evaluation

We used a confusion matrix as a model evaluation metric to find the best and effective model among all models we tested. Here confusion matrix is shown as a 2×2 matrix, having $N = 2$ where two classes are being predicted as approve or disapprove. It results in showing the number of predicted values with the actual values. By continuous test on the datasets, the best model is selected in terms of the score of its classification report.

Performance of each classification algorithm is tested by the measure of its classification report function supported by sklearn [20]. Accuracy is the correct predictions made by the predictive model, whereas precision is determined on the numerical proportion of all predictions that we made with our predictive model are true in nature. Recall gives us an idea about total predictions correctly identified, while F1-score represents the harmonic mean of precision and recall [8]. All actual and predicted values are calculated, and thus metric scores are depicted for each model. These evaluation metrics scores help to determine our final model which is reliable for making predictions.

On the basis of the confusion matrix scores for each model, the evaluation metrics are calculated. Table 2 describes scores of confusion matrix for models.

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}) \tag{2}$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \tag{3}$$

Table 2 Results of confusion matrix for each model

Algorithm	True positive	True negative	False positive	False negative
Logistic regression	21	93	1	28
Decision tree	30	74	20	20
SVM	2	90	4	48

Table 3 Evaluation scores of model

Algorithm	Accuracy score	Precision	Recall	F1-score
Logistic regression	0.79	0.83	0.80	0.77
Decision tree	0.72	0.72	0.72	0.72
SVM	0.64	0.54	0.64	0.63

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \tag{4}$$

$$\text{Score} = (2 * \text{Recall} * \text{Precision})/(\text{Recall} + \text{Precision}) \tag{5}$$

Table 3 shows the resultant values of metrics for each of the algorithms. Based on metrics score, the best scoring model is selected among them to be used as predictive model.

After continuous test using various combinations of predictors, we found the performance of the logistic regression model has accuracy over 79% with a recall of 80%. After the final evaluation of all the models, we see that the accuracy of the logistic regression model is higher compared to other algorithms used. Hence, we can use the logistic regression model as the final predictive model.

5 Conclusion and Future Work

We find that the accuracy of the logistic regression model has a prediction accuracy of nearly 80% which is more than the performance of other models. Further, we can use this model as a prediction model to test with the data of the applicants. The model can provide fast, reliable approach in decision making which can be an alternative to the current procedures adopted by banks for processing loan approval of an applicant.

We can change the prediction variables used in the training of the model for gaining better accuracy and performance of the model. Further, maximum performance in predictability can be achieved through machine learning tools by tuning the variables and implementation with other classification algorithms.

References

1. J. Zurada, Data mining techniques in predicting default rates on customer loans. Databases Inf. Syst. II, 285–296 (2002). https://doi.org/10.1007/978-94-015-9978-8_22
2. R.E. Turkson, E.Y. Baagyere, G.E. Wenya, A machine learning approach for predicting bank credit worthiness, in *2016 Third International Conference on Artificial Intelligence and Pattern Recognition (AIPR)*. (2016). https://doi.org/10.1109/icaipr.2016.7585216
3. A. Vaidya, Predictive and probabilistic approach using logistic regression: application to prediction of loan approval, in *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (2017). https://doi.org/10.1109/icccnt.2017.8203946
4. A.J. Hamid, T.M. Ahmed, Developing prediction model of loan risk in banks using data mining. Mach. Learn. Appl. Int. J. **3**(1), 1–9 (2016). https://doi.org/10.5121/mlaij.2016.3101
5. M.S. Sivasree, T. Rekha Sunny, Loan credibility prediction system based on decision tree algorithm. Int. J. Eng. Res. Technol. **V4**(09) (2015). https://doi.org/10.17577/ijertv4is090708
6. K. Arun, G. Ishan, K. Sanmeet, Loan approval prediction based on machine learning approach. IOSR J. Comput. Eng. **18**(3), 18–21 (2016)
7. K.-S. Shin, T.S. Lee, H.-J. Kim, An application of support vector machines in bankruptcy prediction model. Expert Syst. Appl. **28**(1), 127–135 (2005). https://doi.org/10.1016/j.eswa.2004.08.009
8. B.P. Salmon, W. Kleynhans, C.P. Schwegmann, J.C. Olivier, Proper comparison among methods using a confusion matrix, in *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* (2015). https://doi.org/10.1109/igarss.2015.7326461
9. S. Panigrahi, B. Palkar, Comparative analysis on classification algorithms of auto-insurance fraud detection based on feature selection algorithms. Int. J. Comput. Sci. Eng. **6**(9), 72–77 (2018). https://doi.org/10.26438/ijcse/v6i9.7277
10. P.M. Soni, V. Paul, A novel optimized classifier for the loan repayment capability prediction system, in *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)* (2019). https://doi.org/10.1109/iccmc.2019.8819772
11. G. Arutjothi, C. Senthamarai, Prediction of loan status in commercial bank using machine learning classifier, in *2017 International Conference on Intelligent Sustainable Systems (ICISS)* (2017). https://doi.org/10.1109/iss1.2017.8389442
12. L.T. Priyanka, N. Baby, Classification approach based customer prediction analysis for loan preferences of customers. Int. J. Comput. Appl. **67**(8), 27–31 (2013). https://doi.org/10.5120/11416-6752
13. G. Sudhamathy, Credit risk analysis and prediction modelling of bank loans using R. Int. J. Eng. Technol. **8**(5), 1954–1966 (2016). https://doi.org/10.21817/ijet/2016/v8i5/160805414
14. I.O. Eweoya, A.A. Adebiyi, A.A. Azeta, A.E. Azeta, Fraud prediction in bank loan administration using decision tree. J. Phys: Conf. Ser. **1299**, 012037 (2019). https://doi.org/10.1088/1742-6596/1299/1/012037
15. X. Jency, V.P. Sumathi, J. Sri, An exploratory data analysis for loan prediction based on nature of the clients. Int. J. Recent Technol. Eng. **7**, 176–179 (2019)
16. K.R. Rawate, P.A. Tijare, Review on prediction system for bank loan credibility. Int. J. Adv. Eng. Res. Dev. **4**(12), 860–867 (2017)
17. K.U. Priya, S. Pushpa, K. Kalaivani, A. Sartiha, Exploratory analysis on prediction of loan privilege for customers using random forest. Int. J. Eng. Technol. **7**(2.21), 339 (2018). https://doi.org/10.14419/ijet.v7i2.21.12399
18. N. Torvekar, P.S. Game, Predictive analysis of credit score for credit card defaulters. Int. J. Recent Technol. Eng. **7**, 283–286 (2019)
19. P.D. Singh, A. Chug, Software defect prediction analysis using machine learning algorithms, in *2017 7th International Conference on Cloud Computing, Data Science & Engineering—Confluence* (2017). https://doi.org/10.1109/confluence.2017.7943255
20. Scikit-learn Machine Learning in Python. https://scikit-learn.org/stable