Loan Credibility Prediction using Naïve Bayes, Random Forest, and Logistic Regression: A Comparative Study

Ranyah Taha¹, Sara Alshakrani¹ and Ahmed M. Zeki²

¹College of Science, University of Bahrain
²Department of Informatics Systems, College of IT, University of Bahrain

Abstract—Data mining is a rapidly developing business application field. All banks consider the loan approval process as one of the most significant processes. This research paper conducts a comparative study between three major classifiers, namely: Naïve Bayes, Random Forest, and Logistic Regression in the field of loan approval and credibility. The methodology used in this research is CRISP-DM. This research compares the three classifiers in terms of efficiency to see which one can be used to predict whether a loan applicant would be accepted or denied to support the decision making process. The highest classifier approach in terms of accuracy was Logistic Regression, which had an accuracy rate of 80.1%, as compared to Naïve Bayes and Random Forest that both showed accuracy of 78.9%. Other measures were taken to back up the conclusion using Weka as an analysis tool to support the results and findings.

Index Terms—Naïve Bayes, Random Forest, CRISP-DM, Data Mining, Loan Approval.

I. INTRODUCTION

Owing to extreme rivalry, banks are now fighting it out to achieve an edge over one another in order to boost overall business. Banks also recognized that consumer satisfaction and fraud detection must be a policy method for fair competition (Sivasree & Sunny, 2015). Most banks' primary business is loan distribution. The bulk of a bank's revenue comes from loans provided to customers. These financial institutions charge interest on loans that are sold to consumers. Banks' primary goal is to spend their money in secure customers. Many banks have been processing loans based on a backwards method of authentication and confirmation. However, no bank can promise that a customer selected for a loan application is secure or not.

To prevent this scenario, a Data Mining (DM) approach is used to help in deciding whether to approve a loan application or not for authorizing bank loans. A loan prediction framework is a software that decides whether or not a customer is capable of repaying a loan. Such systems explore a number of considerations, including the customer's marital status, revenue, and spending (Murthy et al., 2020). This method is used by a large number of consumers who have access to a qualified data collection. A necessary model is generated by taking these variables into account. This model is used to generate the necessary output from the test data collection. The outcome would be in the form of a Yes/No answer, “Yes” means that a borrower is capable of repaying the loan, while “No” indicates that the customer is unable to repay the loan. The system will accept loans for consumers based on these criteria (Vangaveeti et al., 2020). This model is applied on a dataset from Kaggle (2019). Three algorithms are used for prediction. They are: Naïve Bayes (NB), Random Forest (RF), and Logistic Regression (LR).

This research would follow the CRISP-DM (Cross Industry Standard Process for Data Mining) framework. The banking mechanism for approving the loan and understanding the 5 C's (character, capacity, condition, capital, and collateral) used by banking officers to accept or deny the loan were both considered in (Tariq, 2019) for the first stage. The aim of this research is to compare the above three methods in loan forecasting. The most accurate model that has the fewest incorrect classified instances will then be chosen.

The rest of this research paper is organized as follows: section 2 will review some of the recent literatures. Section 3 will describe the methodology that is followed in this research including the dataset and dataset preparation. Section 4 will present and discuss the research findings. The paper will then end with a conclusion and future directions.

II. LITERATURE REVIEW

The literature reflects the utilization of DM techniques and how these tools can be used in loan forecasting. To predict whether a consumer defaulted or paid off his or her debt, Zurada (2002) used three different DM methods: Neural Networks (NN), Decision Trees (DT), and LR, as well as an ensemble
model that incorporates the three techniques. He then compared the efficiency of each strategy and examined the probability of default that each loan and community of loans entails and the models were impressive at separating good loans from poor loans.

Riasi & Wang (2016) research uncovered several intriguing findings about the effectiveness of various DM strategies in predicting whether a loan applicant would be accepted or denied. Among the methods studied LAD Tree was ranked first in the overall classifier rankings, followed by Rotation Forest, Logit Boost, Random Forest, and AD Tree. The Logit Boost and AD Tree classifiers, which are included in the LAD Tree classifier, are among the top 5 classifiers in the overall list, which is an interesting finding. Rotation Forest (Accuracy = 85.99%), LAD Tree (Accuracy = 85.58%), and AD Tree (Accuracy = 85.47%) have the highest Accuracy Rates. IB3 achieved the lowest Accuracy Rate (Accuracy = 69.51%).

A comparative study on models used to predict loan risks was conducted in 2018 by Vimala & Sharmili. Using a combination of NB and Support Vector Machine (SVM). NB elegantly combined these methods. Because of its simplicity and robustness, it is commonly used for classification purposes. It classifies data using probability theory. SVM is a type of learning system algorithm that is used to improve the accuracy of classification. They used SVM to increase the accuracy and speed of NB. Their results and analysis showed that the proposed method’s correctness and productivity have been increased, allowing for further extension.

Khan et al. (2021) reported the accuracy of the statistical models based on LR, DT, and RF as 80.94%, 93.64%, and 83.38%, respectively, while with cross-validation the results were 80.94%, 72.21%, and 80.13%. This showed that while the precision of a DT-based model is the best for their dataset, RF was better at generalization, even though its cross validity wasn’t much higher than logistic regression.

Goyal & Kaur (2016) reported that ensemble model outperforms the individual models in terms of estimation. Their model further improves the model's efficiency and precision. They compared multiple models and select the right one for their data using the ensemble model, which assists the company in making the best choice for the customer’s loan request. In terms of precision, the LR model performed marginally better than the radial basis function model, according to the analysis.

A study conducted in 2016 by Murthy et al., three algorithms used by the researchers: K-Nearest Neighbor (KNN), DT, and RF for the approval of the loan prediction to predict loan status. The method produced very accurate results since they used the RF algorithm (Murthy et al., 2020). Table I summarizes the works reviewed earlier.

### III. METHODOLOGY

The Cross-Industry Process for Data Mining (CRISP-DM) (Sivasree & Sunny, 2015) (refer to Figure 1), which is well-known for developing DM ventures, serves as our reference model.

DOI: http://doi.org/10.24086/cocos2022/paper.729

---

**TABLE I**

<table>
<thead>
<tr>
<th>Ref</th>
<th>Methods</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Zurada, 2002)</td>
<td>DT, NN, and LR</td>
<td>The three models were impressive at separating good loans from poor loans</td>
</tr>
<tr>
<td>(Riasi &amp; Wang, 2016)</td>
<td>LAD Tree, Rotation Forest, Logit Boost, RF, and AD Tree</td>
<td>Rotation Forest 85.99%, LAD Tree 85.58%, and AD Tree 85.47% have the highest Accuracy Rates</td>
</tr>
<tr>
<td>(Vimala &amp; Sharmili, 2018)</td>
<td>NB and SVM</td>
<td>Combination of NB and SVM has increased the productivity</td>
</tr>
<tr>
<td>(Khan et al. 2021)</td>
<td>LR, DT, and RF</td>
<td>LR 80.94%, DT 93.64%, and RF 83.38%, RF was better at generalization</td>
</tr>
<tr>
<td>(Goyal &amp; Kaur, 2016)</td>
<td>LR and the radial basis function model</td>
<td>The LR model scored higher precision than the radial basis function model</td>
</tr>
<tr>
<td>(Murthy et al., 2020)</td>
<td>KNN, DT, and RF</td>
<td>RF algorithm presented very accurate results</td>
</tr>
</tbody>
</table>

---

**A. Business Understanding**

This process deals with the objectives of the whole project, as well as the understanding of the requirements.

Both bank staff and applicants benefit from loan prediction. The aim of this paper is to provide a simple, practical, and effective approach for selecting deserving candidates. To do so, DM techniques are used to select the most accurate model with the fewest instances incorrectly classified.

---

**B. Data Understanding**

For implementation purpose, a Kaggle dataset is used which contains data about loan prediction. The dataset comprises of 615 instances and 11 attributes with their respected class as illustrated in Table II.
TABLE II
DESCRIPTION OF THE DATASET

<table>
<thead>
<tr>
<th>No.</th>
<th>Variable Name</th>
<th>Variable Description</th>
<th>Data Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Gender</td>
<td>Male/Female</td>
<td>Char</td>
</tr>
<tr>
<td>2</td>
<td>Married</td>
<td>Applicant married (Y/N)</td>
<td>Char</td>
</tr>
<tr>
<td>3</td>
<td>Dependents</td>
<td>Number of dependents</td>
<td>Int</td>
</tr>
<tr>
<td>4</td>
<td>Education</td>
<td>Graduate/Undergraduate</td>
<td>String</td>
</tr>
<tr>
<td>5</td>
<td>Self_Employed</td>
<td>Self Employed (Y/N)</td>
<td>Char</td>
</tr>
<tr>
<td>6</td>
<td>Income</td>
<td>Applicant income +</td>
<td>Int</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Coapplicant income</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Loan_Amount</td>
<td>Loan amount in thousands</td>
<td>Int</td>
</tr>
<tr>
<td>8</td>
<td>Loan_Amount_Term</td>
<td>Term of loan in months</td>
<td>Int</td>
</tr>
<tr>
<td>9</td>
<td>Credit_History</td>
<td>Credit history meets guidelines</td>
<td>Int</td>
</tr>
<tr>
<td>10</td>
<td>Property_Area</td>
<td>Urban/Semi Urban/Rural</td>
<td>String</td>
</tr>
<tr>
<td>11</td>
<td>Loan_Status</td>
<td>Loan Approved (Y/N)</td>
<td>String</td>
</tr>
</tbody>
</table>

Mostly numeric data, derived from several mathematical derivations, resulting in the data discussed later, which prompts a classification problem. Weka 3.8.5 is used for data analysis.

C. Data Preparation

The dataset has some missing values. To deal with that, the rows with missing values were deleted as they will lead to inaccurate results. Hence, the number of instances decreased from 615 instances to be only 489 instances. The authors applied the Nominal-to-Binary filter that changes the attributes type to be unified type for all nominal attributes to be numeric, and the number of attributes increased from 11 to 15 attributes.

Since the translated data is easier to explore and imagine, the nominal attributes were converted to binary. ‘Applicant income' and 'Coapplicant income' were combined into a single variable called 'income.' From an ethical and privacy point of view, most of the factors are sensitive information which need to be considered when analyzing such data.

D. Modeling

At this point, three methods are used to predict the status of the loan:

- Naïve Bayes (NB): NB has been selected because it assists in the creation of a model that has predictive capability, resulting in a new way of interpreting results. Since the Bayes theorem predicts autonomous speculation, NB is a fundamental methodology based on it. It is commonly used for classification purposes due to its elegant simplicity and robustness. It classifies data using probability theory. NB model is simple to construct and does not require complicated refutation parameters (Hamid & Ahmed, 2016). This classifier is applied using 10-fold cross validation.

- Random Forest (RF): This method helps to predict the loan status in many research papers, and it has been implemented using 10-fold cross validation as well. Because RF is a characterization (and relapse) group learning method that works by constructing a large number of DTs over time and yielding the class that is the mode of the groups yielded by individual trees. The RF algorithm is based on DT law. The distinction is that the DT algorithm only considers one element, while the RF algorithm compares a large number of DTs and returns a result that satisfies the plurality of DTs (Tejaswini et al, 2020).

- Logistic Regression (LR): LR method was implemented using 10-fold cross validation because the linear model is numerically indistinguishable from other regression analyses; its applicability to a wide range of qualitative and quantitative variables is limited (Arun et al, 2016). Also, most of the researchers used this method due to its simplicity and high accuracy.

E. Evaluation

The CRISP methodology's fifth step is the evaluation. Because of the importance and sensitivity of the company target, the correctly and incorrectly classified instances were addressed. As a result, the models' sensitivity, specificity accuracy, precision, recall, F-measure, and Kappa statistic are compared to assess their results. These will be calculated using the equations 1-7 where the number of negative tuples correctly classified by the classifier is referred to as True Negative (TN), the number of negative tuples that were mistakenly labelled as positive is referred to as False Positive (FP), the number of positive tuples that were wrongly labelled as negative is referred to as False Negative (FN), and the term True Positive (TP) refers to positive tuples that has been properly labelled correctly as positive (Keramati et al, 2016). Moreover, the Kappa statistic is implemented where the observed accuracy ($P_0$) and the expected accuracy ($P_e$).

\[
Sensitivity = \frac{TP}{TP + FN} \quad (1)
\]

\[
Specificity = \frac{TN}{TN + FP} \quad (2)
\]

\[
Accuracy = \frac{TN + TP}{TN + FN + TP + FP} \quad (3)
\]

\[
Precision = \frac{TP}{TP + FP} \quad (4)
\]

\[
Recall = \frac{TP}{FN + TP} \quad (5)
\]

\[
F - Measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (6)
\]

\[
Kappa statistic = \frac{P_0 - P_e}{1 - P_e} \quad (7)
\]

The dataset used in this research is unbalanced as shown in Figure 2 where the Loan_Status class has 336 “Yes” as accepted loans, and only 153 “No” as rejected loan applicants.

The Receiver Operating Characteristic (ROC) curve and the Precision-Recall Curve (PRC) are two useful statistical method for explaining classifier results. Furthermore, the Area Under the Curve (AUC) is one of the most often used metrics for assessing a predictive model's accuracy. The AUC was described by Gigliarano et al. (2014) as "the integrated TP rate
over all FP rate values" (Keramati et al, 2016). While the PRC summarizes the trade-off between the TP rate and the positive predictive value. The evaluation is addressed in the results section.

![Dataset class visualization](image)

*Fig. 2. Dataset class visualization.*

**F. Deployment**

Following the completion of the evaluation, the DMs outcome is deployed in the final step, which will also be concluded within the results part.

**IV. RESULTS AND DISCUSSION**

Table III represents the output of the three conducted methods and reported their accuracy where NB model has accuracy of 78.9%, precision of 78.4%, specificity of 49.6%, and sensitivity of 92.2%. The Area under ROC is 76.4% as plotted in Figure 3, and the PRC area is 77.9%.

**TABLE III**

<table>
<thead>
<tr>
<th>Evaluation Indicators</th>
<th>NB</th>
<th>LR</th>
<th>RF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time taking for model building (sec)</td>
<td>0.02</td>
<td>0.01</td>
<td>0.11</td>
</tr>
<tr>
<td>Correctly classified instances</td>
<td>386</td>
<td>392</td>
<td>386</td>
</tr>
<tr>
<td>Incorrectly classified Instances</td>
<td>103</td>
<td>97</td>
<td>103</td>
</tr>
<tr>
<td>Kappa statistic</td>
<td>0.4612</td>
<td>0.3029</td>
<td>0.4526</td>
</tr>
<tr>
<td>Recall</td>
<td>0.789</td>
<td>0.802</td>
<td>0.789</td>
</tr>
<tr>
<td>F-Measure</td>
<td>0.776</td>
<td>0.778</td>
<td>0.773</td>
</tr>
<tr>
<td>Accuracy (%)</td>
<td>78.9%</td>
<td>80.1%</td>
<td>78.9%</td>
</tr>
<tr>
<td>Precision (%)</td>
<td>78.4%</td>
<td>80.1%</td>
<td>78.6%</td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td>92.2%</td>
<td>97.3%</td>
<td>93.4%</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>49.6%</td>
<td>42.4%</td>
<td>47.3%</td>
</tr>
<tr>
<td>ROC Area (%)</td>
<td>76.4%</td>
<td>74.2%</td>
<td>74.8%</td>
</tr>
<tr>
<td>PRC Area (%)</td>
<td>77.9%</td>
<td>76.5%</td>
<td>78.8%</td>
</tr>
</tbody>
</table>

The LR method obtained accuracy of 80.1% and the same result 80.1% for precision. Moreover, the result obtained for the specificity is 42.4% and the sensitivity is 97.3%. Figure 4 illustrates the ROC area which is 74.2% while the PRC area is 76.5%.

The RF method preformed as good as the NB method, where both obtained the same accuracy of 78.9%. The precision is close too where it is 78.6%. The sensitivity for RF 93.4% which is lower than for LR but higher than for NB. The specificity for RF is 47.3%, the ROC area is 74.8%, and the PRC area is 78.8% as shown in Figure 5.

![ROC of the NB method](image)

*Fig. 3. ROC of the NB method.*

![ROC of the LR method](image)

*Fig. 4. ROC of the LR method.*

![ROC of the RF method](image)

*Fig. 5. ROC of the RF method.*

Based on Figures 3, 4, and 5, it can be realized that in terms of ROC and PRC there are no significant changes among the three models as they showed very close results. While in terms of the rest of the measures, LR performed better.

DOI: http://doi.org/10.24086/cocos2022/paper.729
In this analysis, researchers used a cross-validation approach with k=10 fold to boost model evaluation with the goal of reliably predicting loan with more precision and practical and realistic values. In this step, all of tuples in the dataset are used to train and test the methods.

Since the aim was to find the most reliable method with the fewest incorrectly classified instances, the LR model succeeded with the highest percentage of correctly categorized instances and the lowest percentage of wrongly classified instances, also obtained the highest accuracy of 80.1%. The best results of evaluation indicator are shown in Figure 6.

![EVALUATION INDICATOR](image)

**Fig. 6.** Number of correctly and incorrectly classified cases in the three methods.

The findings of this research and the review of the related work demonstrated that using NB and SVM combination improved efficiency of NB model (Vimala & Sharmili, 2018). In this article, however, LR achieved 80.1% accuracy and RF achieved 78.9%. Another research showed that LR Accuracy was 80.945%, and RF accuracy was higher at 83.38% (Khan et al, 2021). While the accuracy of LR at this research paper is 80.1% and this may be attributed to the fact that various datasets were used in the implementation. Another research found that LR was effective at distinguishing between accepted and rejected loans (Zurada, 2002).

**V. CONCLUSION AND FUTURE DIRECTIONS**

This research has suggested a systematic analysis and model creation for loan prediction. The problem of a high ratio of bad loans is extremely important in the financial sector, especially in microfinance banks in various developing and developed countries. Three separate DM techniques were used for the proposed model construction during the experimental process, and their outputs were tested on various parameters. Because of its significant characteristic regarding loan prediction, the best approach was chosen were LR as explained, and recommended based on these criteria.

The current research was limited by the use of small bank’s dataset and the accuracy obtained was lower than expected because of the unbalanced dataset. We might, for future analysis, look at the deployment effects using various approaches and suggest certain strategies to get more accuracy of loan forecast.

**REFERENCES**


DOI: http://doi.org/10.24086/cocos2022/paper.729