Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 7, S. 98-103, 9, 2023

Araştırma Makalesi

IJANSER
https://as-proceeding.com/index.php/ijanser
ISSN: 2980-0811

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 7, pp. 98-103, 9, 2023

Research Article

Comparing Machine Learning Classification Models on a Loan Approval Prediction Dataset

Ladislav Végh [*1], Krisztina Czakóová [1] and Ondrej Takáč [1]

[1] *Department of Informatics, J. Selye University, Slovakia*

[*]*(veghl@ujs.sk) Email of the corresponding author*

(Received: 20 September 2023, Accepted: 05 October 2023)

(3rd International Conference on Innovative Academic Studies ICIAS 2023, September 26-28, 2023)

ATIF/REFERENCE: Végh, L., Czakóová, K. & Takáč, O. (2023). Comparing Machine Learning Classification Models on a Loan Approval Prediction Dataset. *International Journal of Advanced Natural Sciences and Engineering Researches*, 7(9), 98-103.

Abstract – In the last decade, we have observed the usage of artificial intelligence algorithms and machine learning models in industry, education, healthcare, entertainment, and several other areas. In this paper, we focus on using machine learning algorithms in the loan approval process of financial institutions. First, we briefly review some prior research papers that dealt with loan approval predictions using machine learning models. Next, we analyze the loan approval prediction dataset we downloaded from Kaggle, which was used in this paper to compare several machine learning classification models. During this analysis, we observed that credit scores and loan terms are the attributes that probably most affect the result. Next, we divided the dataset into a training set (80%) and a test set (20%). We trained 27 various machine learning models in MATLAB. Three models were optimized with Bayesian optimization to find the best hyperparameters with minimum error. We used 5-fold cross-validation for the validations to prevent overfitting during the training. In the following step, we used the test set on trained models to measure the models' accuracy on unseen data. The result showed that the best accuracy both on validation and test data, more than 98%, was reached with neural networks and ensemble classification models.

Keywords – Machine Learning, Classification, Loan Approval Prediction, Dataset Analysis, Neural Network, Ensemble Model

I. Introduction

Machine learning models and artificial intelligence algorithms can be used in several areas of industry [1], education [2]–[4], healthcare [5]–[6], entertainment [7], and other fields. This article focuses on using machine learning models in financial institutions for loan approval predictions. Even though some threats might arise when financial institutions use artificial intelligence [8], when these modern techniques are used circumspectly, they can significantly decrease the time of some processes, e.g., the time for the decision of loan approvals.

II. Literature Survey

In the literature, we can find several papers that dealt with the topic of loan approvals using machine learning algorithms.

S. S. Sai et al. [9] used three classification techniques: random forest classifier, decision tree, and logistic regression to develop a system for predicting loan status. After scoring the predictions, they reached a result of 79.86%.

In research conducted by M. A. Sheikh et al. [10], a logistic regression model was used on a dataset containing 1500 cases and 10 numerical and 8 categorical attributes to predict loan approvals. The obtained accuracy was 81.1%. A similar result was reached by Y. Divate et al. [11] using a support vector machine algorithm on the same dataset.

N. Pandey et al. [12] compared four classification algorithms: logistic regression, decision tree, support vector machine, and random forest to predict loan approvals. The support vector machine model reached the most accurate result, 79.67%.

In a similar research, A. Shinde et al. [13] used a logistic regression model on over 600 samples to predict loan status. The maximum accuracy of the model was about 82%.

III. Materials and Method

The dataset we used for loan prediction was downloaded from Kaggle [14]. The dataset was first time published on Kaggle in July 2023. The dataset contained 4269 observations and 9 numerical and 3 categorical attributes, including the target variable.

For data analysis and to train and test the classification models, we used MATLAB R2023a.

A. Dataset analysis

First, we examined the target variable of the dataset. As shown in Fig. 1, 62% of loan applications were approved, and 38% were rejected.

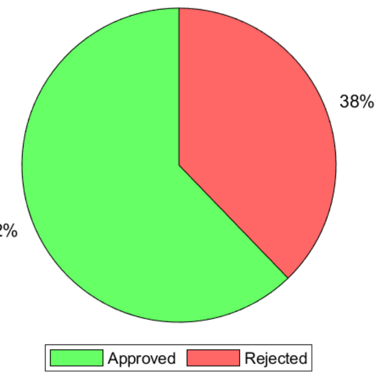

Fig. 1 Percentage of approved and rejected loan applications

Next, we observed all the attributes that might affect the result of the loan approval process.

Fig. 2 shows the distribution of number of dependents of applicants. As we can see on the chart, for the low number of dependents, the ratio of approved and rejected applications is similar to those with higher dependents. We can see a slight difference only for applications with 5 dependents (fewer applications were approved in this category).

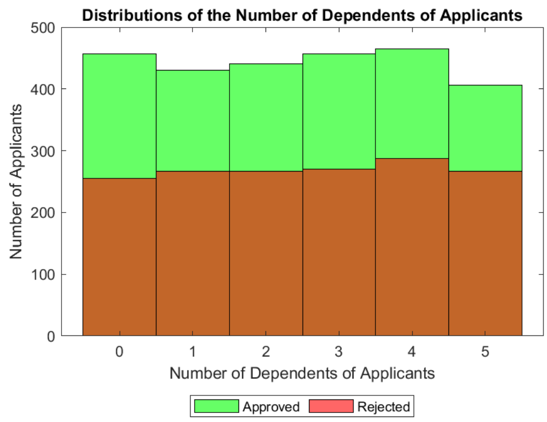

Fig. 2 Distribution of the number of dependents of applicants

Fig. 3 shows the distribution of applicants' education and the distribution of applicants' employment. We cannot observe differences between graduates and not graduates, nor between self-employed and non-self-employed applicants. This might mean that education and employment do not affect the result of the loan approvals.

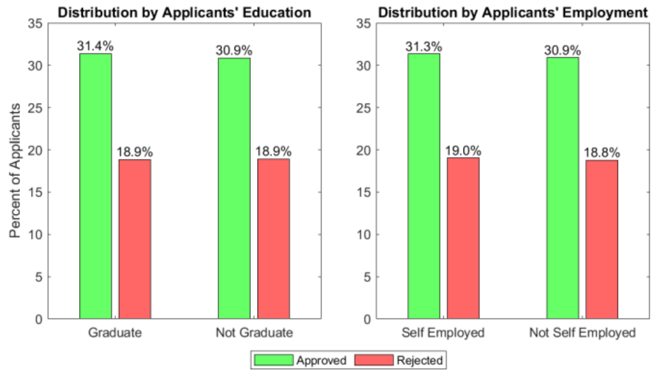

Fig. 3 Distribution of data by applicants' education and employment

In Fig. 4, we can examine applicants' annual incomes and loan amount distribution. We can see slight differences in the ratio of approved and rejected applications for different annual incomes (left chart of the figure). However, we cannot see any clear pattern. We can also observe that for the very high loan amount (last bars in the right chart), about half of the applications were rejected and half approved, while for other loan amounts, more applications were approved than rejected.

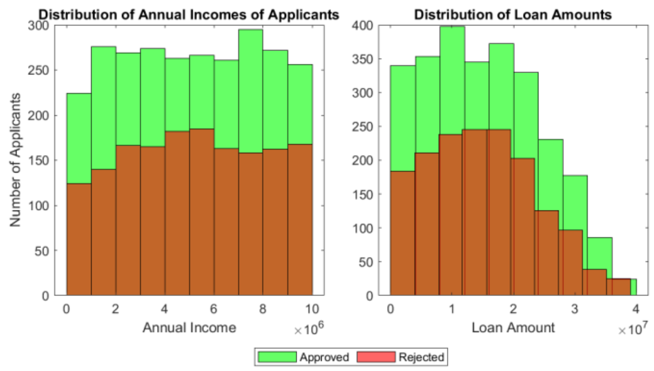

Fig. 4 Distribution of annual incomes and loan amounts

Fig. 5 shows the distribution of loan terms. The chart shows an obvious pattern: applications with short loan terms (less than 6 years) are more often approved and less often rejected than applications with longer loan terms. This means that the loan terms might affect the results of the loan approvals. Financial institutions most likely approve loan applications with shorter loan terms than longer ones.

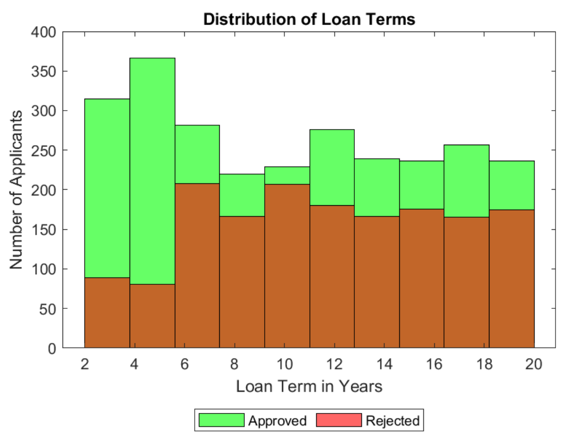

Fig. 5 Distribution of loan terms

In Fig. 6, we can examine the distribution of credit scores for approved applications (left chart) and rejected applications (right chart). We can see that most of the loan applications were rejected with low credit scores (<550), while almost every application was accepted with high credit scores (>600). This observation means that the credit score is the first attribute that financial institutions check in loan applications, which could significantly affect the result of the loan approvals.

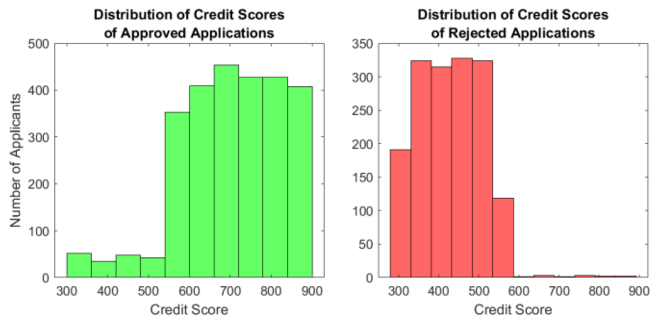

Fig. 6 Distribution of credit scores

The following charts (Fig. 7 and Fig. 8) show the distribution of applicants' residential, commercial, luxury assets and bank asset values. We cannot observe any considerable pattern of approved and rejected applications in these bar charts.

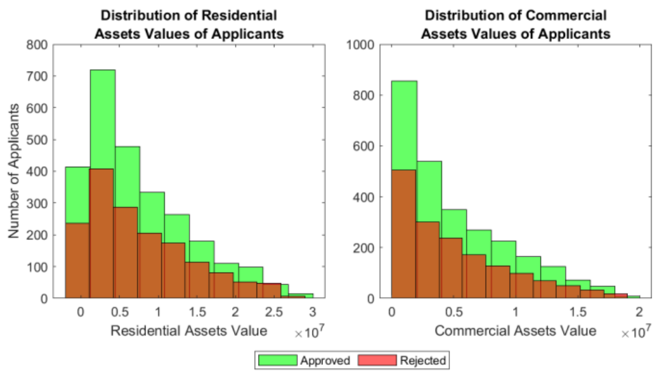

Fig. 7 Distribution of residential assets and commercial assets values of applicants

Fig. 8 Distribution of luxury assets and bank asset values of applicants

B. *Machine Learning Classification Models*

Before the training, we divided the dataset into a training set (80% of data, i.e., 3416 observations) and a test set (20% of data, i.e., 853 observations). We trained 27 machine learning classification models in MATLAB R2023a using the training set. The hyperparameters of three of the models were optimized by Bayesian optimization. We used 5-

fold cross-validation during the training of the models to prevent overfitting. Finally, we used the test set to measure the accuracy of the models on unseen data.

IV. RESULTS

The result of the research is summarized in Table 1. We can see that the best result, 98.45% accuracy on the training set (validation) and 98.83% on the test set, was reached using a narrow neural network. Next, an optimized ensemble classification model achieved 98.42% accuracy on the training set (validation) and 98.83% on the test set. By observing the whole table, we can also notice that the best accuracies were reached with neural networks and ensemble models, followed by the tree, SVM, and Naive Bayes classification models. In contrast, the worst accuracies were obtained with logistic regression and kernel classification models.

Table 1. Validation accuracy, validation total cost, test accuracy, and test total cost of the compared classification models

No.	Model Type	Preset	Accuracy (Validation)	Total Cost (Validation)	Accuracy (Test)	Total Cost (Test)
1	Neural Network	Narrow Neural Network	98.45%	53	98.83%	10
2	Ensemble	Custom Ensemble *	98.42%	54	98.83%	10
3	Ensemble	Boosted Trees	98.33%	57	98.48%	13
4	Neural Network	Custom Neural Network *	98.24%	60	98.01%	17
5	Ensemble	Bagged Trees	97.98%	69	98.48%	13
6	Tree	Fine Tree	97.86%	73	98.01%	17
7	Tree	Custom Tree *	97.86%	73	98.01%	17
8	Neural Network	Trilayered Neural Network	97.72%	78	98.12%	16
9	Ensemble	RUSBoosted Trees	97.60%	82	98.12%	16
10	Neural Network	Bilayered Neural Network	97.37%	90	98.01%	17
11	Tree	Medium Tree	97.25%	94	96.95%	26
12	Neural Network	Medium Neural Network	96.34%	125	96.60%	29
13	Tree	Coarse Tree	96.31%	126	96.25%	32
14	Neural Network	Wide Neural Network	95.99%	137	96.95%	26
15	SVM	Quadratic SVM	94.79%	178	96.48%	30
16	Naive Bayes	Kernel Naive Bayes	94.53%	187	94.26%	49
17	SVM	Cubic SVM	94.47%	189	96.25%	32
18	SVM	Medium Gaussian SVM	93.79%	212	93.90%	52
19	Naive Bayes	Gaussian Naive Bayes	93.30%	229	93.43%	56
20	SVM	Linear SVM	92.89%	243	91.91%	69
21	SVM	Coarse Gaussian SVM	92.77%	247	92.03%	68
22	Binary GLM Logistic Regression	Binary GLM Logistic Regression	92.04%	272	91.68%	71
23	SVM	Fine Gaussian SVM	82.06%	613	82.77%	147
24	Efficient Logistic Regression	Efficient Logistic Regression	62.21%	1291	62.25%	322
25	Efficient Linear SVM	Efficient Linear SVM	62.21%	1291	62.25%	322
26	Kernel	SVM Kernel	60.66%	1344	59.44%	346
27	Kernel	Logistic Regression Kernel	59.84%	1372	61.20%	331

* Bayesian optimization was used to optimize the model hyperparameters.

Table 2 shows the hyperparameters of the best model. As we can see, one fully connected layer was used in the narrow neural network, the first layer size was 10, and the ReLU activation function was used.

Table 2. Hyperparameters of model no. 1 (neural network)

Hyperparameter	Value
Preset:	Narrow Neural Network
Number of fully connected layers:	1
First layer size:	10
Activation:	ReLU
Iteration limit:	1000
Regularization strength (Lambda):	0
Standardize data:	Yes

Fig. 9 shows the validation confusion matrix of the narrow neural network. Most cases were correctly predicted. However, there were 26 false negative (0.76%) and 27 false positive predictions (0.79%).

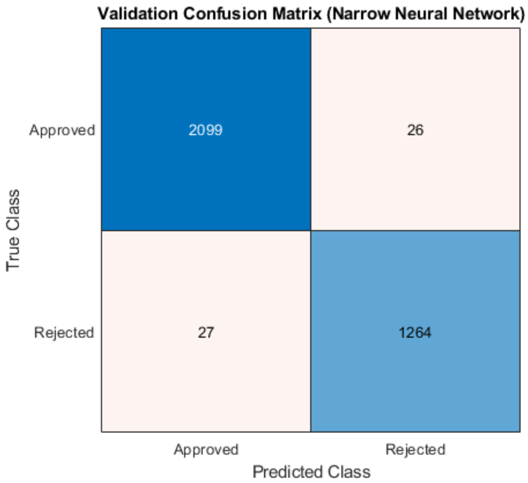

Fig. 9 Validation confusion matrix of model no. 1 (neural network)

The second model in Table 1 that reached 98.42% validation accuracy and 98.83% test accuracy was an optimized ensemble model. We used Bayesian optimization to find the model's best hyperparameters. Fig. 10 shows the minimum classification error plot. We can observe that the classification error decreased to 0.015809 during the optimization.

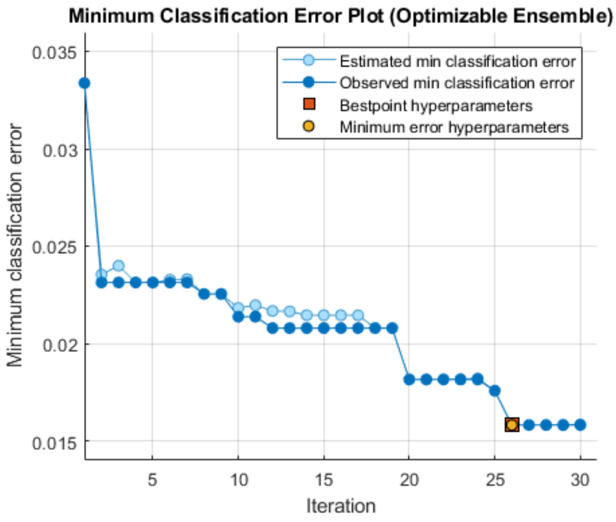

Fig. 10 Minimum classification error plot of model no. 2 (ensemble) optimization

The optimized ensemble model's hyperparameters are shown in Table 3. The decision tree learner and GentleBoost ensemble method reached the best hyperparameters.

Table 3. Minimum error hyperparameters (also bestpoint hyperparameters) of model no. 2 (ensemble)

Hyperparameter	Value
Preset:	Optimizable Ensemble
Learner type:	Decision tree
Ensemble method:	GentleBoost
Number of learners:	12
Learning rate:	0.057721
Maximum number of splits:	223

Fig. 11 shows the validation confusion matrix of the optimized ensemble model. Most cases were correctly predicted. However, there were 20 false negative (0.59%) and 34 false positive predictions (1,00%).

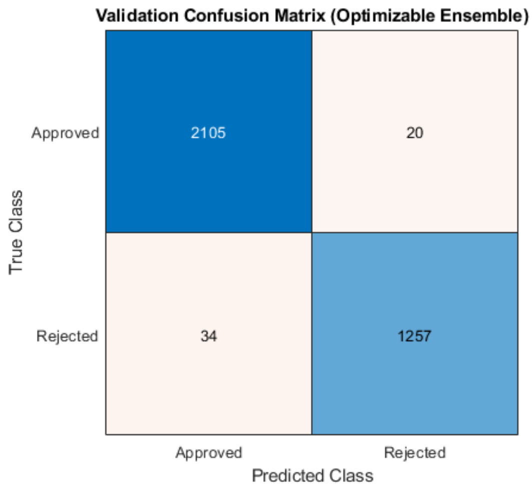

Fig. 11 Validation confusion matrix of model no. 2 (ensemble)

V. Conclusion

After a short introduction and brief literature review, this paper examined a loan approval prediction dataset. Next, we compared the accuracy of 27 machine learning classification models on the given dataset using MATLAB. The results showed that the best accuracies were reached with neural networks and ensembled machine learning models. The outcomes of this paper could help create similar models for loan approval predictions for financial institutions. Using appropriate machine learning classification models, the length of the process of loan approvals can be significantly reduced.

References

[1] M. T. Fülöp, M. Gubán, Á. Gubán, and M. Avornicului, "Application Research of Soft Computing Based on Machine Learning Production Scheduling," *Processes 2022*, vol. 10, issue 3, paper 520, March 2022. https://doi.org/10.3390/pr10030520

[2] N. Annuš, "Usability of Artificial Intelligence to Create Predictive Models in Education," in *EDULEARN23 Proceedings*, 2023, pp. 5061–5065. https://doi.org/10.21125/edulearn.2023.1328

[3] N. Annuš, "Weigh the Pros and Cons of Using Artificial Intelligence in Education," *International Journal of Science, Engineering and Technology (IJSET)*, vol. 11, issue 3, May 2023.

[4] J. Udvaros and N. Forman, "Artificial Intelligence and Education 4.0," in *INTED2023 Proceedings*, 2023, pp. 6309–6317. https://doi.org/10.21125/inted.2023.1670

[5] A. Al Kuwaiti, K. Nazer, A. Al-Reedy, S. Al-Shehri, A. Al-Muhanna, A. V. Subbarayalu, D. Al Muhanna, F. A. Al-Muhanna, "A Review of the Role of Artificial Intelligence in Healthcare," *Journal of Personalized Medicine*, vol. 13, issue 6, paper 951, June 2023. https://doi.org/10.3390/jpm13060951

[6] A. Zahlan, R. P. Ranjan, D. Hayes, "Artificial intelligence innovation in healthcare: Literature review, exploratory analysis, and future research," *Technology in Society*, vol. 74, paper 102321, August 2023. https://doi.org/10.1016/j.techsoc.2023.102321

[7] S. Gupta. (2023) AI in Entertainment. [Online]. Available: https://www.scaler.com/topics/ai-in-entertainment/

[8] Z. Illési and V. Honfi, "A Security Assessment of AI, Related to the Financial Institutions," in *Security-Related Advanced Technologies in Critical Infrastructure Protection*, 2022, pp. 85–94. https://doi.org/10.1007/978-94-024-2174-3_7

[9] S. S. Sai, A. B. Krishna, P. Ganthi, S. K. Kankanala, S. Tinnaluri, and V. Simhadri, "Machine Learning Algorithms for Predicting the Loan Status," *Journal of Survey in Fisheries Sciences*, vol. 10, no. 25, pp. 3677–3685, 2023.

[10] M. A. Sheikh, A. K. Goel, and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," in *Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020)*, IEEE, 2020, pp. 490–494.

[11] Y. Divate, P. Rana, and P. Chavan, "Loan Approval Prediction Using Machine Learning," *International Research Journal of Engineering and Technology (IRJET)*, vol. 8, issue 5, pp. 1741–1745, May 2021.

[12] N. Pandey, R. Gupta, S. Uniyal, and V. Kumar, "Loan Approval Prediction using Machine Learning Algorithms Approach," *International Journal of Innovative Research in Technology (IJIRT)*, vol. 8, issue 1, pp. 898–902, June 2021.

[13] A. Shinde, Y. Patil, I. Kotian, A. Shinde, and R. Gulwani, "Loan Prediction System Using Machine Learning," in *International Conference on Automation, Computing and Communication 2022 (ICACC-2022)*, ITM Web of Conferences 44, 2022, paper 03019, p. 4. https://doi.org/10.1051/itmconf/20224403019

[14] A. Sharma. (2023) Loan-Approval-Prediction-Dataset. Loan Approval Dataset used for Prediction Models. [Online]. Available: https://www.kaggle.com/datasets/architsharma01/loan-approval-prediction-dataset