Using Random Forest as a Novel Approach to Loan Prediction and Comparing Accuracy to the Support Vector Machine Algorithm

K. Sravani [1], Mahaveerakannan.R [2*]

[1]Research Scholar, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India, Pincode: 602105

[2*]Project Guide, Corresponding Author, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu, India, Pincode: 602105

ABSTRACT

Aim: The main aim of the research is to predict loans using RF (Random Forest) over a novel Support Vector Machine Algorithm (SVM). **Materials and Methods:** Random Forest and a novel Support Vector Machine are implemented in this research work. Sample size is calculated using G power software and determined as 10 per group with a pretest power of 80%, a threshold of 0.05% and a confidence interval of 95%. **Result:** Random Forest provides a higher score of 85.30% compared to the Novel Support Vector Machine algorithm with 75.10% in predicting loans. There is a significant difference between the two groups, with a insignificance value of 1.000 ($p > 0.05$). **Conclusion:** The results show that the Random Forest algorithm for loan prediction appears to generate better accuracy than the Support Vector Machine

Keywords: Novel Support Vector Machine Algorithm, Loan Prediction, Machine Learning, Random Forest, Credit Score

INTRODUCTION

The work deals with the prediction of new loans using the Random Forest algorithm on the Novel Support Vector Machine algorithm. Prediction using machine learning successfully compared Random Forest with the Novel Support Vector Machine algorithm. The banking industry always needs a more accurate predictive modeling system for many problems. Predicting defaults is a difficult task for the banking sector. Loan status is one of the indicators of loan quality. It doesn't show everything right away, but it's a first step in the loan process. Loan status is used to build a credit score model (Khadse 2021). The credit scoring model is used for an accurate analysis of credit data to find insolvent and valid customers. The goal of this document is to create a credit score model for your credit data. Various machine learning techniques are used to develop the financial credit score model (Maheswari and Narayana 2020). In this article, we propose an analytics model based on a machine learning classifier for credit data. We use the combination of Min-Max normalization and Support Vector Machine. The goal is implemented using the software package tool Colab. This proposed model provides the necessary information with the utmost precision. It is used to predict the status of loans in commercial banks using a machine learning classifier (Sujatha et al. 2021).(Venu and Appavu 2021; Gudipaneni et al. 2020; Sivasamy, Venugopal, and Espinoza-González 2020;

Sathish et al. 2020; Reddy et al. 2020; Sathish and Karthick 2020; Benin et al. 2020; Nalini, Selvaraj, and Kumar 2020)

For financial institutions and the banking sector, it is very important to have predictive models for their financial activities, as they play an important role in risk management. Predicting loan defaults is one of the critical issues they focus on, as a huge loss of revenue could be avoided by predicting the customer's ability to pay on time (Demraoui, Eddamiri, and Hachad 2022). In this paper, several classification methods (Naïve Bayes, Decision Tree, and Random Forest) are used for prediction. Several comprehensive preprocessing techniques are applied to the dataset, and three different feature extraction algorithms are used to improve accuracy and performance (Sunitha and Adilakshmi 2021). The results are compared using the F1 precision measurement, and an improvement of more than 3% was achieved (Gupta et al. 2020). The study aims to improve by incorporating Random Forest and comparing performance with that of Support Vector Machine. The proposed model improves classifiers to achieve greater accuracy in loan prediction (Gupta et al. 2020; Sheikh, Goel, and Kumar 2020).Previously our team has a rich experience in working on various research projects across multiple disciplines(Venu and Appavu 2021; Gudipaneni et al. 2020; Sivasamy, Venugopal, and Espinoza-González 2020; Sathish et al. 2020; Reddy et al. 2020; Sathish and Karthick 2020; Benin et al. 2020; Nalini, Selvaraj, and Kumar 2020).The existing research study that has the least accuracy for predicting loan approval is 73.12% using machine learning algorithms (Rembart and Soliman 2017) and this will create a research gap to predict the loan approval for a customer. Hence, the aim of the research work is to use the Novel Logistic Regression technique in order to classify the deserving applicants and reject the undeserved applicants by achieving better accuracy.

MATERIALS AND METHODS

This study setting was done in the Soft Computing Laboratory, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences. The number of required samples in the research is two, in which group 1 is Random Forest compared with group 2 of Novel Support Vector Machine Algorithm. The samples were taken from the device and iterated 10 times to get the desired accuracy with G power of 80%, threshold of 0.05%, and confidence interval of 95%. A dataset consisting of a collection of data was downloaded from Kaggle.

Random Forest

Random forests, or random decision forests, are an ensemble learning method for classification, regression, and other tasks that operate by constructing a multitude of decision trees at training time. For classification tasks, the random forest produces the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random forests generally outperform decision trees, but their accuracy is lower than gradient-boosted trees.

Pseudocode for Random Forest

```
from sklearn.ensemble import RandomForestClassifier
RF = RandomForestClassifier()
RF.fit (X_train_scaled, y_train)
RF_pred = RF.predict(X_test_scaled)
classification_report (y_test, RF_pred))
RF_SC = accuracy_score(RF_pred,y_test)
```

```
print('Accuracy Score of Random Forest Classifier: ', accuracy_score(y_test, RF_pred))
matrix4=confusion_matrix(y_test, RF_pred)
```

The Novel Support Vector Machine Algorithm

Support Vector Machine is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. Support Vector Machine is an implementation of gradient boosted decision trees designed for speed and performance.

Pseudocode for Novel Support Vector Machine Algorithms

```
from sklearn.svm import SVC
svm = SVC()
SVM.FIT (X_train_scaled, y_train)
svm_pred = svm.predict(X_test_scaled)
classification_report (y_test, SVM_pred))
SVM_SC = accuracy_score(svm_pred,y_test)
print('Accuracy Score of Super Vector Machine: ', accuracy_score(y_test, svm_pred))
matrix3=confusion_matrix(y_test, svm_pred)
```

Recall that the testing setup includes both hardware and software configuration choices. The laptop has an Intel Core i5 5th generation CPU with 12GB of RAM, an x86-based processor, a 64-bit operating system, and a hard drive. Currently, the software runs on Windows 10 and is programmed in Python. Once the program is finished, the accuracy value will appear. Procedure: Wi-Fi laptop connected. Chrome to Google Collaboratory search Write the code in Python. Run the code. To save the file, upload it to the disc, and create a folder for it. Log in using the ID from the message. Run the code to output the accuracy and graph.

Statistical Analysis

SPSS is a software tool used for statistical analysis. The proposed system utilized 10 iterations for each group, with predicted accuracy noted and analyzed. An independent sample t-test was done to obtain significance between the two groups. Loan opening and closing price parameters are independent variables, and loan prediction is a dependent variable (Singh et al. 2021).

RESULTS

Table 1 shows the accuracy value of iteration of Random Forest and Support Vector Machine. Table 2 represents the group statistics results credit score, which depicts Random Forest with a mean accuracy of 78.581% and the standard deviation is 3.028. The Support Vector Machine has a mean accuracy of 71.773% and a standard deviation of 3.028. The proposed Random Forest algorithm provides better performance compared to the Novel Support Vector Machine algorithm. Table 3 shows the independent sample T-test value for Random Forest and Support Vector Machine with a mean difference of 8.1 and a standard error difference of 0.67. The insignificance value is observed as 1.000 ($p > 0.05$).

Table 1. Group Statistics of Random Forest with Support Vector Machine by grouping the iterations with a sample size of 10, minimum = 65, maximum = 79, and Standard Derivation = 0.513. Comparison between RF and SVM algorithms with N = 10 samples of the dataset with the highest accuracy of 85.30% and 75.10% in sample 1 (when N = 1) using the dataset size of 914 and 70% of training and 30% of testing data.

Table 2. Group Statistics of Random Forest with Support Vector Machine by grouping the iterations with a sample size of 10, mean value of 78.581, and standard deviation of 3.028. The RF has got 3.028 standard deviations with a 0.957 standard error, while another SVM algorithm recorded 2.058 standard deviations with a 0.627 standard error. Also, the independent sample t-test was utilized to analyze the precision of two calculations, and a measurably massive contrast was taken note of, with the RF model obtaining a mean of 78.581% accuracy and the SVM having a mean of 71.773% accuracy.

Table 3. Independent Sample Test of Accuracy and Precision, Calculate P-value = 0.001, insignificant value = 1.000, mean difference = 5.000, and confidence interval = (1.354-2.155). The Random Forest and Support Vector Machine are significantly different from each other.

Figure 1 shows the bar graph comparison of the mean of accuracy on Random Forest and Novel Support Vector Machine algorithms. The mean accuracy of Random Forest is 78.581% and Support Vector Machine is 71.773%.

DISCUSSION

In this study, loan prediction using the Random Forest algorithm has significantly greater accuracy, approximately 78.581% compared to the Support Vector Machine (71.773%). Random Forest appears to produce more consistent results with a minimal standard deviation.

Similar results in the paper were 79% accurate with Support Vector Machine, which was used to predict the loan. The reported Support Vector Machine proposed work has a 74% accuracy, which is used to predict stock market and loan performance (Aditya Sobika et al. 2021). Work proposed by Random Forest has 79% better accuracy. According to his research, SVM is a metric for measuring loans that is used with both traditional and modern methods. Random Forest has the highest accuracy, credit score, and Support Vector Machine will achieve the lowest accuracy compared to other machine learning techniques, which varies by 3% compared to other machine learning techniques (Karthikeyan and Ravikumar 2021). When you use Support Vector Machine to predict the loan, you will have key issues to fake in this document, which shows that Random Forest has a minimum accuracy of 78%. Increasing the data set value tends to achieve the desired accuracy, credit score. Random Forest works best with a combination of other machine learning algorithms (Udaya Bhanu and Narayana 2021).

The limitation of this search is that it cannot provide appropriate results for small data. In this model, it is not possible to consider all the parameters of the characteristic variables given for training. The future purpose of the proposed work will be classification-based loan prediction using class labels for less time complexity.

CONCLUSION

In this study, the results show that the Random Forest algorithm for loan prediction appears to generate better accuracy than the Support Vector Machine. The results obtained show that the RF algorithm has found 85.30% more accuracy on the provided dataset than the SVMn Algorithm (75.10%). Table 3 has a calculated two-tailed significant value of p 0.001. The results corresponding to equal variances are considered for analysis.

DECLARATION

Conflict of Interests

No conflict of interests in this manuscript

Authors Contribution

Author KS was involved in data collection, data analysis, manuscript writing. Author MVK was involved in conceptualization, data validation, and critical review of manuscript.

Acknowledgement

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding: We thank the following organizations for providing financial support that enabled us to complete the study.

1. Inoaura Technologies, Chennai.
2. Saveetha University.
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering

REFERENCES

1. Aditya Sobika, K., M. Jamuna, H. Limitha, Mn Saroja, and Sn Shivappriya. 2021. “An Empirical Study on Loan Prediction Using Logistic Regression and Decision Tree.” In *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*. IEEE. https://doi.org/10.1109/icaeca52838.2021.9675505.
2. Benin, S. R., S. Kannan, Renjin J. Bright, and A. Jacob Moses. 2020. “A Review on Mechanical Characterization of Polymer Matrix Composites & Its Effects Reinforced with Various Natural Fibres.” *Materials Today: Proceedings* 33 (January): 798–805.
3. Demraoui, Lamiae, Siham Eddamiri, and Lamiae Hachad. 2022. “Digital Transformation and Costumers Services in Emerging Countries: Loan Prediction Modeling in Modern Banking Transactions.” In *AI and IoT for Sustainable Development in Emerging Countries*, 627–42. Cham: Springer International Publishing.
4. Gudipaneni, Ravi Kumar, Mohammad Khursheed Alam, Santosh R. Patil, and Mohmed Isaqali Karobari. 2020. “Measurement of the Maximum Occlusal Bite Force and Its Relation to the Caries Spectrum of First Permanent Molars in Early Permanent Dentition.” *The Journal of Clinical Pediatric Dentistry* 44 (6): 423–28.
5. Gupta, Anshika, Vinay Pant, Sudhanshu Kumar, and Pravesh Kumar Bansal. 2020. “Bank Loan Prediction System Using Machine Learning.” In *2020 9th International Conference System Modeling and Advancement in Research Trends (SMART)*. IEEE. https://doi.org/10.1109/smart50582.2020.9336801.
6. Karthikeyan, and Pushpa Ravikumar. 2021. “A Comparative Analysis of Feature Selection for Loan Prediction Model.” *International Journal of Computers & Applications* 174 (11): 49–55.
7. Khadse, Dr Kavita. 2021. “APPLICATIONS OF MACHINE LEARNING IN LOAN PREDICTION SYSTEMS.” *Linguistica Antverpiensia* 2021 (3): 3658–74.
8. Maheswari, P., and C. H. V. Narayana.

2020. “Predictions of Loan Defaulter - A Data Science Perspective.” In *2020 5th International Conference on Computing, Communication and Security (ICCCS)*. IEEE. https://doi.org/10.1109/icccs49678.2020.9277458.

9. Nalini, Devarajan, Jayaraman Selvaraj, and Ganesan Senthil Kumar. 2020. “Herbal Nutraceuticals: Safe and Potent Therapeutics to Battle Tumor Hypoxia.” *Journal of Cancer Research and Clinical Oncology* 146 (1): 1–18.
10. Reddy, Poornima, Jogikalmat Krithikadatta, Valarmathi Srinivasan, Sandhya Raghu, and Natanasabapathy Velumurugan. 2020. “Dental Caries Profile and Associated Risk Factors Among Adolescent School Children in an Urban South-Indian City.” *Oral Health & Preventive Dentistry* 18 (1): 379–86.
11. Rembart, Franz, and Erfan Soliman. 2017. “Loan Demand in Jordanian Microfinance Market: Interest Rate Elasticity and Loan-Acceptance Prediction via Logistic Regression.” *Enterprise Development and Microfinance*. https://doi.org/10.3362/1755-1986.16-00009.
12. Sathish, T., and S. Karthick. 2020. “Gravity Die Casting Based Analysis of Aluminum Alloy with AC4B Nano-Composite.” *Materials Today: Proceedings* 33 (January): 2555–58.
13. Sathish, T., D. Bala Subramanian, R. Saravanan, and V. Dhinakaran. 2020. “Experimental Investigation of Temperature Variation on Flat Plate Collector by Using Silicon Carbide as a Nanofluid.” In *PROCEEDINGS OF INTERNATIONAL CONFERENCE ON RECENT TRENDS IN MECHANICAL AND MATERIALS ENGINEERING: ICRTMME 2019*. AIP Publishing. https://doi.org/10.1063/5.0024965.
14. Sheikh, Mohammad Ahmad, Amit Kumar Goel, and Tapas Kumar. 2020. “An Approach for Prediction of Loan Approval Using Machine Learning Algorithm.” In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*. IEEE. https://doi.org/10.1109/icesc48915.2020.9155614.
15. Singh, Vishal, Ayushman Yadav, Rajat Awasthi, and Guide N. Partheeban. 2021. “Prediction of Modernized Loan Approval System Based on Machine Learning Approach.” In *2021 International Conference on Intelligent Technologies (CONIT)*. IEEE. https://doi.org/10.1109/conit51480.2021.9498475.
16. Sivasamy, Ramesh, Potu Venugopal, and Rodrigo Espinoza-González. 2020. “Structure, Electronic Structure, Optical and Magnetic Studies of Double Perovskite Gd2MnFeO6 Nanoparticles: First Principle and Experimental Studies.” *Materials Today Communications* 25 (December): 101603.
17. Sujatha, C. N., Abhishek Gudipalli, Bh Pushyami, N. Karthik, and B. N. Sanjana. 2021. “Loan Prediction Using Machine Learning and Its Deployement on Web Application.” In *2021 Innovations in Power and Advanced Computing Technologies (i-PACT)*. IEEE. https://doi.org/10.1109/i-pact52855.2021.9696448.
18. Sunitha, M., and T. Adilakshmi. 2021. “Addressing Long Tail Problem in

Music Recommendation Systems.” *International Journal of Computational Systems Engineering*. https://doi.org/10.1504/ijcsyse.2021.10045577.

19. Udaya Bhanu, L., and Dr S. Narayana. 2021. “Customer Loan Prediction Using Supervised Learning Technique.” *Mathematical Sciences Research Journal. An International Journal of Rapid Publication* 11 (6): 403–7.
20. Venu, Harish, and Prabhu Appavu. 2021. “Experimental Studies on the Influence of Zirconium Nanoparticle on Biodiesel–diesel Fuel Blend in CI Engine.” *International Journal of Ambient Energy* 42 (14): 1588–94.

TABLES AND FIGURES

Table 1. Comparison between RF and SVM algorithms with N = 10 samples of the dataset with the highest accuracy of 85.30% and 75.10% in sample 1 (when N = 1) using the dataset size of 914 and 70% of training and 30% of testing data.

Sample (N)	dataset size	RF Accuracy in %	SVM Accuracy in %
1	914	85.30	75.10
2	872	82.78	74.85
3	811	81.21	73.16
4	772	78.50	72.50
5	715	77.54	72.10
6	697	77.10	71.12
7	667	76.94	70.80
8	617	76.44	70.50
9	578	75.80	69.10
10	548	74.20	68.50

Table 2. Group Statistics of Random Forest with Support Vector Machine by grouping the iterations with Sample size 10, Mean value of RF 78.581 and SVM 71.773%, Standard Derivation = 3.028. Descriptive An Independent Sample Test of Accuracy and Precision is applied to the dataset in SPSS. Here it specifies equal variances with and without assuming a T-Test Score of two groups with each sample size of 10.

Group Statistics					
Accuracy	Groups	N	Mean	Std deviation	Std. Error Mean
	Random Forest	10	78.581	3.028	0.957
	The Support Vector Machine	10	71.773	2.058	0.627

Table 3. Independent Sample Test of Accuracy and Precision, Calculate P-value = 0.001, insignificant value = 1.000, mean difference = 5.000, and confidence interval = (1.354-2.155). The Random Forest and Support Vector Machine are significantly different from each other.

Accuracy	Independent Samples Test								
	Levene's Test for Equality of Variances					T-test for Equality of Means			
	F	Sig	t	df	Sig (2-tailed)	Mean Difference	Std.Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
Equal variance is assumed.	0.000	1.000	3.693	18	0.002	5.000	1.354	2.155	7.845
Equal variances are not assumed			3.693	18.000	0.002	5.000	1.354	2.155	7.845

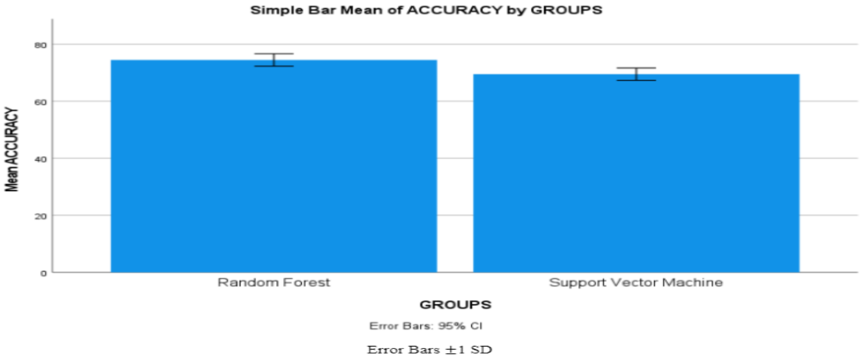

Fig. 1. Bar Graph Comparison on mean accuracy of Random Forest (78.581%) and Support Vector Machine (71.773%). X-axis: Random Forest, Support Vector Machine, Y-axis: Mean Accuracy with ±1 SD