Loan Approval Prediction Based on Tabu Search Optimization Technique

Rishika Chittimalla[1]
CSE(AIML)
Sr University
Warangal,India
chrishika7@gmail.com

Alekhya Cholleti[2]
CSE(AIML)
Sr University
Warangal,India
cholletialekhya@gmail.com

Sanjana Gunda[3]
CSE(AIML)
Sr University
Warangal,India
gundasanjana22@gmail.com

Vyshnavi Manthurthi[4]
CSE(AIML)
Sr University
Warangal,India
vyshnavimanthurthi2004@gmail.com

Anshu Kumar Dwivedi[5]*
CSE
Sr University
Warangal, India
d.anshukumar@sru.edu.in

***Abstract*— The banks and their customers know loan approval methods. Although traditional methods like taking all the relevant information is time-consuming, there is a need for technology, which can result and machine learning algorithms being applied to the data sets that contain the past data of customers to help in predicting whether the loan is being approved or not. Now the question is how much these algorithms are predicting accurate results, our research paper tells how the optimization we applied will maximize the number of loan approvals and minimize the risk analysis. We applied the tabu search optimization along with a few machine learning algorithms on the loan data set, and our accuracy is about 98%.**

Keywords—Classification, logistic regression, decision tree, SVM, Random Forest, Tabu Search Optimization.

I. Introduction

In the financial services area especially in the banking system, the loan approval process is very much important for the economy [1]. There is a need to ensure that the loans are approved wisely as they need to be repaid to maintain the balance of the economy [5]. Various factors such as credit history, income level, loan amount and other factors need to be examined thoroughly before the approval of a loan [7]. By using tabu search optimization there can be an increase in adeptness of the loan approval process [3]. The Tabu search approach is developed for solving problems with many possible solutions or combination problems [3]. It reiterates from one result to another neighbouring result for finding the best solution according to predefined criteria thereby helping in the decision-making loan approval processes [4].In recent times, prediction models have gained a great response in various fields, contributing insights into complicated situations such as predicting loan defaults and improving software testing adeptness. Gazi et al. [1] explored this in finance, by studying different classifiers and feature selection methods for predicting loan approvals. They found that pre-processing methods and algorithms, which involve feature selection, make a huge contribution to improving model accuracy and adeptness. Meanwhile, Amit et al. [2] have focused on the loan approval processes, which highlights insights such as credit scores and income levels. In software testing, Javier et al. [3] introduced a new method, Tabu Search, for making high branch handling tests. In our research paper, the selection and extraction process includes the tabu search optimization where the results are very accurate which helps in maximizing the number of loan approvals and minimizing the risks.

II. Literature survey

According to Gazi et.al. [1], the research investigated the default prediction of how the various classifiers and feature selection methods include the preprocessing techniques using enhanced minority class prediction across classifiers. Feature selection algorithms also improved the model performance, especially when using large datasets. This research tells the crucial role of preprocessing in model accuracy and the importance of feature selection for both the accuracy and adeptness of the models. Future research may explore additional search space of classifiers and feature selection approaches across various dataset sources. According to Amit et.al. [2], the prediction process starts with data preparation, including cleaning, processing, and handling missing values, followed by experimental analysis and model building. The highest accuracy achieved on the original dataset is 81. The study of this paper indicates that applicants with lower credit scores are less likely to receive loan approval due to a higher risk of default. Inversely those applicants with higher income and lower loan amounts requested can mostly be approved, suggesting a correlation with repayment capability. Insights like gender and marital status appear insignificant in the loan approval decision-making. To achieve high branch coverage, Javier et al. [3], proposed the Tabu Search, an adaptive software testing method. A tabu search-based method for hardware-software partitioning was presented by Mehdi et al.

[4], to reduce the logic area. Superior design results in terms of hardware cost are demonstrated by comparison testing against simulated annealing, tabu, genetic algorithms, and a combination method. According to Tejaswini et. al. [5], the paper proposed that risk evaluation for credit is very necessary. It is also one of the main tasks for loan approval for the banking system. This system allows us to forecast whether the candidate is appropriate for the validation process using various machine-learning approaches. This methodology included data collection following the data selection then training and testing on the data set and finally, the analysis of the result using various machine learning models. It is also used for fixing various errors and bugs in programmed predictive systems. According to Pimcharee et. al. [6], the paper compared the data mining techniques with feature selection and without feature techniques for credit approval. Three different machine learning models were used without feature selection and further with feature selection two more models were used, namely, information gain and Chi-square Finally concluded that the best results were without feature selection for credit approval. According to Vishal et. al. [7], the prediction for loan approval is modernized by using machine learning algorithms for two datasets. Here for one data set, based on the two target values the eligibility for the candidate's approval is decided whether he or she can pay the loan back or not. For the second data set, taking various inputs from users provided a forecast of whether the loan could be issued or not. This paper is helpful for banks to minimize losses and increase loan approvals. According to Harun et.al. [8], the Tabu Genetic Algorithm (TGA), hybrid meta-heuristic methods increase adeptness by improving solution quality, and speed by incorporating Tabu Search (TS) into the genetic algorithm. (GA). The researchers are arguing about how these new models help in improving the quality of the data, research that suggests integration of human memory characters into tabu search, this combination of insights from psychology and heuristics used for further optimizing problem-solving strategies. According to Alejandra et.al. [9], this study introduced the Greedy Randomized Adaptive Search Procedure (GRASP) for the K subset problem, which receives a much faster solutions process than traditional methods like logistic regression, KNN, etc. By ranking the research space of searches and implementing response strategies, it aims to improve the quality of solutions without losing future research efficiency. Moreover, strategies like the risk of repetitive greed can increase computation time and guarantee further improvement.

According to Safwan et.al. [10], this study evaluates machine-learning algorithms for loan charge estimates from 1,000 loan data sets, which shows the efficacy of decision support for loans. The results indicate a satisfactory accuracy ratio and the neural network is effective in this task. The Neural Network emphasizes neural networks as a good tool for risk assessment in banks and emphasizes the importance of algorithm selection in decision management for lending.

III. Approach and Methodology

A. Data Pre-processing

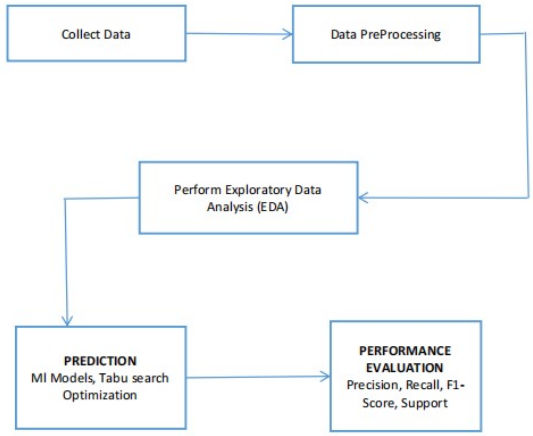

Fig 1. The architecture of loan approval prediction

Features Selection and Extraction

It is a useful method for determining which qualities are compatible with the projections of loan acceptance. Pre-processing the data, specifying the evaluation metrics, initializing the tabu list, specifying the neighbourhood search, objective function, and tabu search algorithm, assessing the completed solution, and interpreting the outcomes are the most popular methods. It is a useful method for determining which qualities are compatible with the projections of loan acceptance. Pre-processing the data, specifying the evaluation metrics, initializing the tabu list, specifying the neighbourhood search, objective function, and tabu search algorithm, assessing the completed solution, and interpreting the outcomes are the most popular methods.

B. Models Construction and Validation

1. **Random Forest:** An approach to collective learning is Random Forest Regression. During training, several decision trees are constructed. The average (for regression) or mode (for classification) of each tree's contribution to the final prediction is used. It is renowned for its capacity to handle different types of data and reliability against overfitting.

2. **Logistic Regression:** Logistic regression is a statistical technique used to model binary outcomes by estimating the likelihood of an event occurring based on one or more predictors. This logistic function maps output to a probability between 0 and 1 whereby coefficients that best fit the data are calculated to predict the probability of an event for new observations.

3. **Support Vector Machine (SVM):** SVR is a regression method that extends Support Vector Machines to issues related to regression. It locates a hyperplane that minimizes the error margin and most accurately matches the data. SVR works well at identifying complex patterns in data and is especially helpful when working with high-dimensional data

4. **K Nearest Neighbour(KNN):** K-Nearest Neighbors (KNN) is a type of non-parametric supervised learning algorithm used for classification and regression tasks. It involves finding, among the training set, K closest data points to a new point, then assigning their majority class (classification) or averaging their values (regression) to the new point. The value of k i.e., the number of closer neighbors considered affects the performance of this model.

5. **Decision Tree:** A decision tree can be described as is that it's a type of supervised learning algorithm most commonly applied within both regression and classification. This would divide feature space into sections based on input features' values and create a tree-like structure having leaf nodes as well as decision nodes. Each node will make some decision concerning impurity minimization in regards to variance for classification/regression cases thus forming sequences involving binary decisions finally ending with estimated classes at leaves.

6. **Tabu Search:** A meta-heuristic approach inspired by human behaviour that is engaged in addressing combination optimization problems. It iterates from one solution to another neighbouring solution as exploring the solution space, but it also keeps track of the past solutions in a short-term memory called tabu list. It is only because of its memory that the tabu search can find the optimal solutions.

Model validation

A Tabu Search-based validation involves a process where a TS generates a neighbourhood solution from the current solution, when the best solution is generated the current solution is updated with the best solution, and the neighbourhood solution is replaced.

When we think of how well this process makes the process of getting the best solution possible, we need to know the actual process behind it:

- Initially, we need to set up an empty tabu list (TL), a current solution c and a best solution bs.
- We need to generate a few neighbourhood solutions C of the current solution c.
- Based on the objective function the BS will be saved as a c'
- The search space checks whether c' is better than bs, if this condition is true then the tabu list will be updated else,
- If solution c' is already on the tabu list the revisited solution will be removed from it.
- If all the termation conditions are satisfied the search will terminate and an optimal solution will be generated.

IV. Result

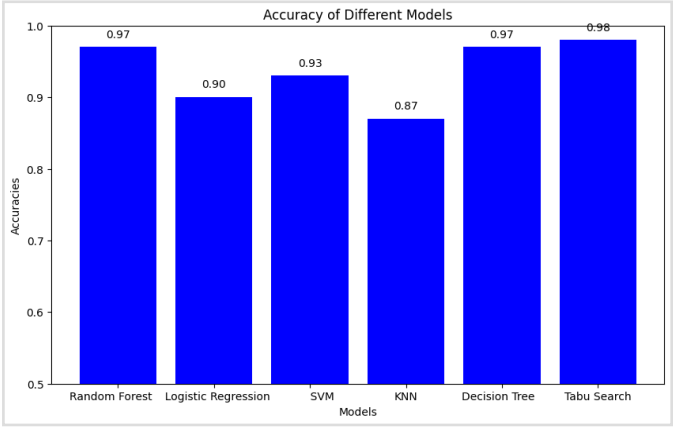

Fig 2 Comparision of accuracies of various models.

Fig 2 compares the accuracies of various machine learning models performed on the data set. The minimum accuracy was found to be 87% for KNN and the maximum accuracy was found to be 98% for the tabu search optimization technique. The performance evaluation of various models applied is recorded as below:

Table I: Classification report for Random forest.

Class	Precision	Recall	F1-Score	Support
0	0.97	0.95	0.96	471
1	0.97	0.98	0.96	810
Accuracy			0.97	1281
Macro Avg	0.97	0.97	0.97	1281
Weighted Avg	0.97	0.97	0.97	1281

Table II: Classification report for Logistic Regression.

Class	Precision	Recall	F1-Score	Support
0	0.87	0.87	0.87	471
1	0.93	0.92	0.92	810
Accuracy			0.90	1281
Macro Avg	0.90	0.90	0.90	1281
Weighted Avg	0.90	0.90	0.90	1281

Table III: Classification report for Support Vector Machine.

Class	Precision	Recall	F1-Score	Support
0	0.89	0.93	0.91	471
1	0.96	0.93	0.94	810
Accuracy			0.93	1281
Macro Avg	0.92	0.93	0.93	1281
Weighted Avg	0.93	0.93	0.93	1281

Table IV: Classification report for K's nearest neighbour.

Class	Precision	Recall	F1-Score	Support
0	0.81	0.86	0.83	471
1	0.92	0.88	0.90	810
Accuracy			0.87	1281
Macro Avg	0.86	0.87	0.87	1281
Weighted Avg	0.88	0.87	0.88	1281

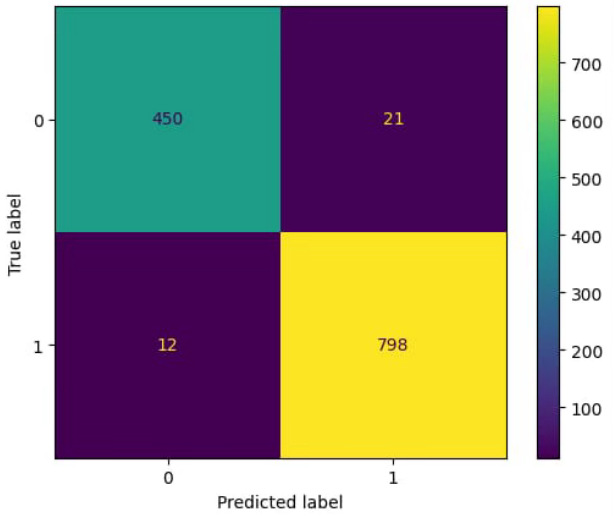

Fig 3. Confusion matrix for Decision Tree.

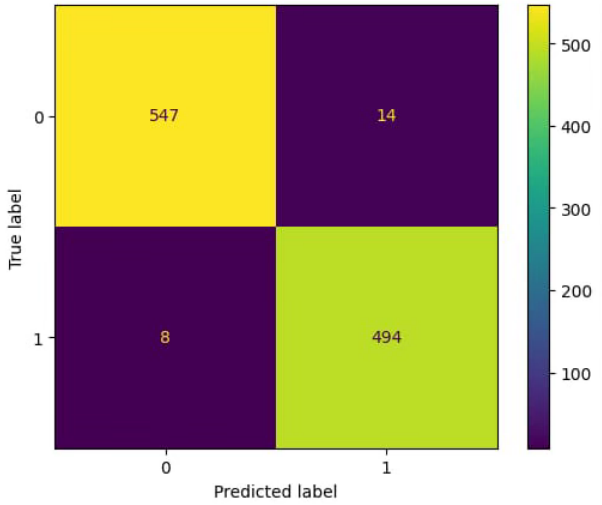

Fig 4. Confusion matrix for Tabu Search.

V. Conclusion

In this paper, we applied the tabu search optimization technique along with various other machine learning techniques including, Random forest, logistic regression, SVM, KNN, and Decision trees for predicting the approval of a loan. The best accuracy was obtained for tabu search optimization with 98% compared to all other machine learning models used. Not only accuracy but it also showed various performance evaluation techniques such as precision, recall, F1-Score, and Support. This optimization technique is used for maximizing the loan approvals and minimizing the risk analysis. Shortly, we intend to perform deep learning and ensemble learning algorithms such as XG Boost, and Ada Boost for predicting the approval of a loan. We even intend to work for processing automated systems which may lead to higher accuracy for loan approval.

References

[1] Al-qerem, A., Al-Naymat, G., & Alhasan, M. (2019). Loan Default Prediction Model Improvement through Comprehensive Preprocessing and Features Selection. In 2019 International Arab Conference on Information Technology (ACIT). IEEE. DOI: [10.1109/ ACIT47155.2019.8994246]

[2] Ahmad Sheikh, M., Goel, A. K., & Kumar, T. (2020). An Approach for Prediction of Loan Approval using Machine Learning Algorithm. (IEEE Xplore Part Number: CFP20V66-ART; ISBN: 978-1-7281-4108-4). Galgotias University, Greater Noida, India.

[3] Díaz, E., Tuya, J., & Blanco, R. (2003). Automated Software Testing Using a Metaheuristic Technique Based on Tabu Search. In Proceedings of the 18th IEEE International Conference on Automated Software Engineering (ASE'03). University of Oviedo, Gijón, Asturias, Spain.

[4] Jemai, M., Dimassi, S., Ouni, B., & Mtibaa, A. (2017). A metaheuristic based on the tabu search for hardware-software partitioning. Retrieved from http://journals.tubitak.gov.tr/elektrik/ [10.3906/. elk-1501-64]

[5] Tejaswini, J., Kavya, T. M., Ramya, R. D. N., Triveni, P. S., & Maddumala, V. R. (2020). Accurate loan approval prediction based on a machine learning approach. Journal of Engineering Science, 11(4), 523-532.

[6] Pimcharee, K., & Surinta, O. (2022). Data Mining Approaches in Personal Loan Approval. Engineering Access, 8(1), 15-21.

[7] Singh, V., Yadav, A., Awasthi, R., & Partheeban, G. N. (2021, June). Prediction of modernized loan approval based ML approach. In 2021 International Conference on Intelligent Technologies (CONIT) (pp. 1-4). IEEE.

[8] Pirim, H., Eksioglu, B., & Bayraktar, E. (2008). Tabu Search: a comparative study.

[9] Casado, A., Pérez-Peló, S., Sánchez-Oro, J., & Duarte, A. (2022). A GRASP algorithm Tabu Search improvement solving the maximum intersection of k-subsets problem. journal of Heuristics, 28(1), 121–146. https://doi.org/ 10.1007/s10732-022-09490-8

[10] Alsaleem, M. Y., & Hasoon, S.O. (2020). Predicting bank loan risks using ML algorithms. Raf. J.Comp. & Math's., 14(1), 149. [10.33899/csmj.2020.164686]

[11] J. R. Quinlan. Improved use of continuous attributes in c4.5.Journal of Artificial Intelligence Research, 4:77-90, 1996. [9] Lee, T. S., Chiu, C. C., Chou, Y. C., & Lu, C. J. (2006). Mining the customer credit using classification regression tree and multivariate adaptive regression splines. Computational Statistics Data Analysis,50, 111

[12] Mattei, Pierre-Alexandre. (2019). A Parsimonious Tour of Bayesian Model Uncertainty.

[13] Salzberg, Steven. (1996).Book Review: "C4.5: Programs for Machine Learning" by J. Ross Quinlan.

[14] Stahl, F., & Jordanov, I. (2012). An overview of the use of neural networks for data mining tasks. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(3), 193–208. doi: 10.1002/widm.1052

[15] Supriya, P., Pavani, M., Saisushma, N., Vimala, N. Vikas, K. (2019). Loan Prediction by using Machine Learning Models. International Journal of Engineering Techniques, 5(22), pp.144-148.

[16] Wang, Di & Wu, Qi & Zhang, Wen. (2019). Neural Learning of Online Consumer Credit Risk.

[17] M. Bentlemsan et al., "Random Forest and Filter Bank Common Spatial Patterns for EEG-Based Motor Imagery Classification", International Conference on Intelligent Systems, Modelling and Simulation, pp. 235 – 238, 2014.

[18] Chioka (2013, Aug, 30), Available: http://www.chioka.in/ classimbalance-problem/. [Accessed: 2- Jan- 2019]

[19] Hong Zhang, Yong-gong Ren and Xue Yang, "Research Text Feature Selection Algorithm Based on Information Gain and Feature Relation Tree", 10th Web Information System and Application Conference, pp. 446 – 449, 2013.

[20] Ho-duck Kim et al., "Genetic Algorithm Based Feature Selection Method Development Pattern Recognition", SICE-ICASE International Joint Conference, pp. 1020 – 1025, 2006.