Grenze International Journal of Engineering and Technology, June Issue

Comparative Analysis of SVM, Logistic Regression, Random Forest, and XGBoost for Credit Risk Assessment

Y Jeevan Nagendra Kumar[1], Siri Chandhana Allenki[2], Likki Sanjana[3], Niveditha Pallempati[4] and Ch. Vidyadhari[5]
[1-5]Gokaraju Rangaraju Institute of Engineering and Technology, Department of Information Technology, Hyderabad, India
Email: jeevannagendra@griet.ac.in, chandhana535@gmail.com, sanjanalikki05@gmail.com, nivedithapallempati@gmail.com, chalasanividyadhari@gmail.com

***Abstract*— With the evolving world of financial technology, accurate credit risk prediction has become crucial to minimize loan defaults and enhance financial inclusion. Traditional credit assessment relies primarily on historical credit data, rendering credit unaffordable for individuals with thin credit files. This research attempts to bridge this gap by conceptualizing and comparing different machine learning algorithms for loan approval prediction. The study contrasts Support Vector Machines (SVM), Logistic Regression, Random Forest, and Extreme Gradient Boosting (XGBoost) performance in the loan applicant classification task with demographic and financial features. The methodology includes data preprocessing, feature engineering, and unsupervised clustering to identify underlying patterns in the data prior to training. This is followed by model training and testing on the basis of key performance measures such as accuracy, precision, recall, and F1-score. By comparing the performance of models in varying conditions, the paper demonstrates the strengths and weaknesses of each approach. The results are intended to build more reliable, data-driven decision systems that enhance the more equitable and faster loan approval process and reduce the risk of defaults.**

***Index Terms*— Credit Risk Prediction, Loan Approval, Machine Learning, Model Comparison, SVM, Logistic Regression, Random Forest, XGBoost, Clustering, Financial Inclusion, Predictive Analytics.**

I. INTRODUCTION

In today's financial environment, proper credit risk analysis is critical for lending organizations to reduce defaults and increase financial inclusion. Historical credit history-based traditional credit assessment techniques primarily use the past credit history, which tends to leave out applicants with poor or no credit history, thus limiting their loan access. Recent advances in machine learning algorithms present promising alternatives by utilizing heterogeneous applicant information to enhance the accuracy and fairness of prediction. This study emphasizes the development and comparison of various machine learning algorithms—Support Vector Machine (SVM), Logistic Regression, Random Forest, and Extreme Gradient Boosting (XGBoost), for loan applicant classification into risk levels. Through demographic and financial attributes, the models should make loan approval decisions better than traditional methods.

Grenze ID: 01.GIJET.12.2.156_20

Before, researchers approached the issue of class imbalance with methods such as GAN oversampling [1] or using heuristic ideas to select features and applied SMOTE [2]. In addition to deep learning being helpful for credit risk and fraud detection [3], the main focus of this study is on simplified and efficient machine learning. By detailed preprocessing, feature engineering, and class imbalance handling, this research assesses the performance of these models through accuracy, precision, recall, and F1-score metrics. The comparative assessment discloses the merits and demerits of each model, giving insights into their applicability to automated loan approval systems. Finally, this research contributes to the development of data-driven models that improve decision-making speed, minimize bias, and provide credit opportunities to a larger society.

II. Related Work

[1] Yanyu Zhuang and Hua Wei have tackled the problem of imbalanced datasets for personal credit risk prediction. Their solution entails the use of a generative adversarial network, say a Wasserstein GAN with gradient penalty (WGAN-GP), to oversample minority classes and a gradient boosting algorithm such as LightGBM to create a stable prediction model. This would involve attempting to enhance the precision and verifiability of credit risk predictions, especially in situations where default cases are much less common than non-default cases in the given dataset. Through the creation of synthetic data points for the less-represented class, the WGAN-GP can be utilized to balance the dataset, enabling the LightGBM model to learn better and produce more precise predictions.

[2] Zeinab Hassani, Mohsen Alambardar Meybodi, and Vahid Hajihashemi researched developing credit risk measurement using learning algorithm-based feature selection. Their Fuzzy Information and Engineering publication investigates approaches to determine the most critical variables that have a bearing on effective credit risk prediction. The research probably explores the application of methods like the Synthetic Minority Over-Sampling Technique (SMOTE), which corrects imbalanced datasets by creating synthetic samples for the minority class, and the Firefly Algorithm, an optimization algorithm possibly utilized for choosing the best subset of features. Through strategically choosing a smaller subset of extremely informative features, the authors seek to create credit risk assessment models that are not only more precise in their forecasts but also more efficient and explainable.

[3] Mienye and Jere's (2024) IEEE Access review reviews the expanding application of deep learning (DL) in credit card fraud detection (CCFD). The authors narrate and compare different DL methodologies such as CNN, RNN, LSTM, and GRU used on this task. The paper further reviews important features like the correct evaluation metrics of performance and frequent difficulties faced in training DL models to detect fraud as well as suggested solutions. By analyzing recent studies and experimental outcomes on a real-world dataset, the paper emphasizes the efficacy of DL architectures in detecting fraudulent transactions and presents useful insights for researchers as well as practitioners in the domain.

[4] In Alejandrino, Bolacoy, and Murcia's (2023) IJECE study investigated data mining techniques for loan default prediction in the Philippines. They applied supervised algorithms like J48, k-NN, Naïve Bayes, and Logistic Regression to a loan dataset. During training, k-NN with k=3 (78.38%) and Logistic Regression (77.31%) achieved the highest classification accuracy. When used to predict defaults on unseen data, k-NN 3 identified 48 non-defaulters and 52 defaulters, while Logistic Regression predicted 44 non-defaulters and 56 defaulters. The research underscores the value of data mining for financial institutions in predicting loan defaults and improving risk management.

[5] Safa Bozkurt Coşkun and Münevver Turanlı compared the performance of CatBoost, XGBoost, and LightGBM on the Kaggle Home Credit Default Risk dataset in their 2023 JAMSI article, performing a credit risk estimation using boosting techniques. The study concluded that all three gradient boosting methods produced similar results, although CatBoost registered a higher AUC score when crediting tendency was considered, whereas XGBoost and LightGBM registered a slight decline. The authors had performed parameter tuning with grid search. The tuned models had a higher tendency to grant credit, improving recall of the repayment class but decreasing precision. Interestingly enough, CatBoost's AUC value increased significantly after tuning but that of LightGBM and XGBoost remained relatively unaltered. The authors saw that LightGBM and XGBoost did well in all cases even in their implementations without any trimmings. But they added that this could have been biased as a result of overfitting the overwhelming non-crediting class in the very imbalanced data.

[6] The study by Kui Wang, Meixuan Li, Jingyi Cheng, Xiaomeng Zhou, and Gang Li focuses on improving personal credit risk evaluation using XGBoost. Recognizing limitations in existing credit characteristics, they proposed a two-step model. First, they employed Logistic Regression for feature selection, comparing standard, AIC, and BIC approaches to identify the most relevant indicators for credit risk. The Logistic Regression model

with the best AUC and accuracy is used to create a credit risk index. Second, they compared the performance of XGBoost, Decision Tree, and KNN for classifying borrowers as likely to default or not, ultimately selecting XGBoost due to its superior performance across AUC, accuracy, and Type II error metrics. The effectiveness of this combined approach is demonstrated using 10,000 real credit data points from X Bank.

[7] Dansana et al.'s 2023 Engineering Reports research employs the Random Forest model to determine the influence of loan attributes on bank loan prediction. In an effort to enhance loan approval and mitigate risk, they studied variables such as gender, education, employment, and loan duration. The Random Forest model indicates that banks should use a variety of customers' characteristics other than wealth in making credit decisions. Visual guidance led the model, which forecasts loan approval by combining decisions from various trees and estimates feature importance.

III. Classification Algorithms in Machine Learning for Credit Risk Prediction

Credit risk prediction requires the determination of the probability that a loan seeker will default or settle the loan. In an attempt to solve this classification problem, machine learning algorithms have been used because they can master sophisticated patterns in past data. Four common classification algorithms were used in this research: Extreme Gradient Boosting (XGBoost), Random Forest, Support Vector Machine (SVM), and Logistic Regression.

A. Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is an efficient, scalable implementation of the gradient boosting algorithm. It creates an ensemble of decision trees sequentially, where each additional tree tries to minimize the error of the preceding ones through gradient descent. This process enhances model performance via regularization, preventing overfitting.

XGBoost is capable of dealing with missing values, efficiently deals with big datasets, and generates feature importance scores; hence, it is suitable for credit scoring scenarios with a blend of numerical and categorical features. Many studies have found that XGBoost outperforms CatBoost and LightGBM when the dataset involved has high dimensionality or is unbalanced [5].

B. Random Forest

Random Forest is an ensemble learner that builds many decision trees when trained. It trains each tree on a different subset of data and votes for the majority class in classification. The random element introduced by sampling the data and features at tree building increases model diversity and lowers overfitting risk. Random Forest is famous for its stability, high-dimensional data handling, and relatively good interpretability, thus forming a strong candidate for credit risk prediction tasks.

C. Support Vector Machine

Support Vector Machine (SVM) is a supervised learning method used to find the best hyperplane that maximally separates various classes in the feature space. Its performance is best seen in high-dimensional situations, where, by using kernel functions, non-linearly separable data is mapped to higher dimensions where linear separation is achieved successfully.

SVMs can deliver high accuracy rates, especially in binary classification problems, but are computationally expensive to handle large datasets and require careful hyperparameter tuning, such as the regularization parameter and the kernel type used. However, SVM is a suitable tool for credit risk assessment, especially in situations where a clean distinction is seen between risky and non-risky applicants.

D. Logistic Regression

Logistic Regression is a straightforward but effective statistical technique applied for binary classification as well. Logistic Regression estimates the chance of an input belonging to a specific class through a logistic function.

Although the model supposes a linear relationship between the input features and the log-odds of the target, logistic regression is recommended because it is easy to understand, efficient, and simple to implement. Logistic regression works effectively when the data is in linear form and is a reliable baseline model for credit scoring tasks.

Apart from using Logistic Regression as a classifier, it has also been used to help select important features for models such as XGBoost in credit risk classification problems [6].

IV. METHODOLOGY

This chapter outlines the methodology used in developing, training, and testing machine learning models for loan approval status prediction. Methodology involves data preprocessing, feature engineering through clustering, handling class imbalance, and training and testing of the models using four different classifiers.

A. Data Collection

The dataset used for this study is similar to real-time data, though the sources remain unknown. The features include applicant income, co-applicant income, loan amount, tenure of loan, credit history, gender, education, employment, marital status, and assets. The target feature of our study was to obtain the loan approval status. To enhance the size and quality, data augmentation procedures were employed. This process includes the generation of new records based on data entries in the dataset. Using data augmentation allowed the model to generalize well, making overfitting less likely in the validation and comparison of models.

B. Data Preparation

Data was imported and initially cleaned by removing unwanted columns such as customerid, to avoid excessive bias or data leakage.

C. Encoding Categorical Variables

The data set contained numerous categorical features that were critical for credit risk scoring (e.g., purpose, marital_status_sex, current_account). These were encoded into numerical form through Label Encoding. Additionally, manual categorical mappings where domain knowledge drove appropriate numerical ordering or grouping were employed, further enhancing model interpretability and performance.

D. Feature Engineering through Clustering

To enhance feature representation, K-Means clustering was performed on selected numeric features related to loan attributes (install_rate, amount, duration). Data was divided into four clusters, and cluster labels were introduced as a new feature. This aids the derivation of concealed data patterns and can improve the efficiency of classification.

E. Data Splitting and Feature Scaling

The target variable credit was label-encoded to transform class labels into a numeric form. Features were scaled with StandardScaler, which standardizes features to a zero mean and unit variance. The dataset was split into training and testing subsets with a 70-30 stratified ratio to preserve class distributions.

F. Class Imbalance with SMOTE

SMOTE was selected to adjust class proportions, as previous studies have confirmed its effectiveness in helping the minority class in financial datasets [2]. SMOTE generates synthetic examples of minority class samples and thus balances the set, enabling models to better detect underrepresented classes. Oversampling was performed to meet a minimum size requirement per class. For experimental reasons, the SMOTE was applied to the test set as well, but usually in real-world implementations, this is avoided.

G. Model Training

During this research, all the machine learning experiments were performed in Jupyter Notebook. It is an extremely popular tool that is used to make it easy for researchers and practitioners to write and execute code. It contains a straightforward user interface for data exploration, chart generation, and model building. This makes it excellent for building and testing predictive models. The dataset was preprocessed and augmented with the Synthetic Minority Over-sampling Technique (SMOTE) before training to deal with class imbalance. This allowed the minority class to generate artificial examples so that the models were provided with balanced and informative training data. Much like k-NN and Logistic Regression in regional studies [4], our study assesses different classifiers and compares their performances. Four classification methods were utilized to develop and test the loan eligibility prediction model: Extreme Gradient Boosting, Random Forest, Support Vector Machine, and Logistic Regression. All models were trained using the augmented training set with the appropriate settings for their respective learning methods.

- XGBoost was chosen because of its scalability, regularization capability, and ability to handle missing values and high-dimensional data efficiently.

- Random Forest is a technique consisting of numerous decision trees. Since Random Forest is easy to interpret and allows you to sort features by importance [7], it was applied to identify main factors that matter in getting a loan approved.
- A Support Vector Machine with an RBF kernel performed well in solving non-linear decision boundaries.
- Logistic Regression was included as a baseline classifier, offering interpretability and simplicity for binary classification tasks.

All the models were preprocessed in the same way to enable comparison of their performance in a fair and meaningful way. We tested them with critical classification metrics such as Accuracy, Precision, Recall, and F1-Score. These metrics provide a balanced perspective of the performance of the models, particularly when there is class imbalance.

We then trained all the models to enhance their performance to be able to handle new data and prevent overfitting. We aimed to find the optimal parameters that reduce the errors of classification to maintain the model stable. This precise training allowed all the algorithms to be optimized to fit the unique features of the dataset to enable correct and accurate predictions regarding the eligibility to acquire loans.

Model Evaluation

The trained models were compared on the test set based on key classification measures: accuracy, precision, recall, and F1-score. Confusion matrices were also calculated for thorough error examination. This comparative evaluation helped us determine the optimal algorithm to predict loan approval based on balanced predictive performance and class imbalance stability.

V. RESULTS

The study tested the performance of four machine learning algorithms—XGBoost, Random Forest, Logistic Regression, and Support Vector Machine (SVM) to predict credit risk. We measured how effective each model was by using accuracy, precision, recall, F1-score, and R^2 scores. We examined the confusion matrices to analyze the results of the classifications.

Model Performance: The metrics used to assess each model are summarized in Table 1.

TABLE I: PERFORMANCE EVALUATION METRICS FOR ML ALGORITHMS

Model	Accuracy	Precision	Recall	F1 Score
XGBoost	0.8627	0.866	0.8627	0.8621
Random Forest	0.8561	0.8593	0.8561	0.8555
Logistic Regression	0.8586	0.8609	0.8586	0.8581
SVM	0.7907	0.7909	0.7907	0.7906

Looking at Table 1, XGBoost obtained the best performance, with an R^2 of 0.7946 and an accuracy score of 86.27%. Random Forest and Logistic Regression both demonstrated similar performance, although Random Forest had slightly worse scores. SVM produced the weakest accuracy of between 79% and 79.5% and the lowest R^2 of 0.6869. Graphs were produced to help illustrate how the model performed compared to previous models.

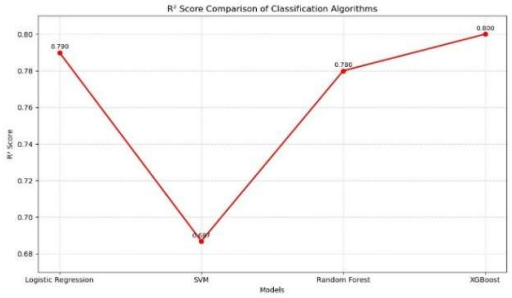

Figure 1. R^2 Score Comparsion of Classification Algorithms

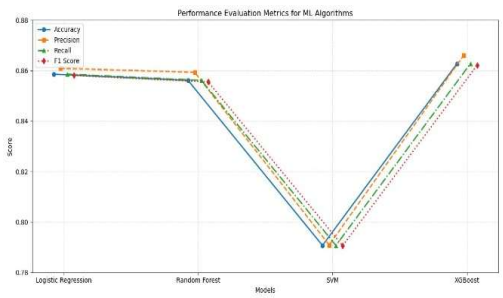

Figure 2. Performance Evaluation Metrics for ML Algorithms

Figure 1 demonstrates the differences in R^2 scores for the four models. Out of all models used, XGBoost displayed the greatest R^2 score of 0.7946, suggesting it is the best at understanding the variance in the dataset. According to the study [5], XGBoost is better than CatBoost and LightGBM when facing financial inputs that

are high-dimensional and are imbalanced. The Random Forest gave a score of 0.7847, while Logistic Regression obtained a score of 0.7884. SVM performed with the lowest R^2 score at 0.6869. In Figure 2, a side-by-side comparison of accuracy, precision, recall, and F1-score for all models is shown. Using XGBoost, the team achieved an accuracy of 86.27% and good precision, recall, and F1-scores all over 86%. The performance of Logistic Regression and Random Forest was similar, each reaching an accuracy of about 85.6%. Based on all metrics, SVM did not perform as well as expected, with scores close to 80%. Being strong and comparable with ensemble methods matches prior observations in [6], where Logistic Regression was applied for classifying and also selecting important features in credit scoring models. This accounts for its strong results in our study, especially since it uses proper preprocessing.

Confusion Matrix: They clearly show how the models classified results correctly or incorrectly. For example: XGBoost made 3,309 true negatives and 2,811 true positives, with a small number of misclassifications (550 false positives and 1,050 false negatives). XGBoost had fewer misclassifications than Random Forest. The number of false positives (1,300) and false negatives (1,139) in SVM suggests it was not very good at classification. The results from the confusion matrix reveal that having unequal class sizes can influence the effectiveness of the model. Models such as SVM, which struggle with class imbalance, often report a lot of mistakes. This is why SMOTE, a form of oversampling, has proved helpful in much research, including [1] and [2], for detecting minority classes in financial data. Because of SMOTE, XGBoost, and Logistic Regression were able to recognize and evaluate classes that make up a small percentage of the data.

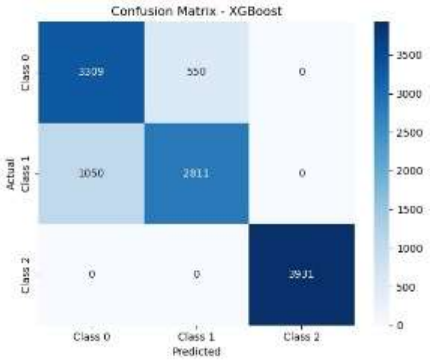

Figure 3. Confusion Matrix – XGBoost

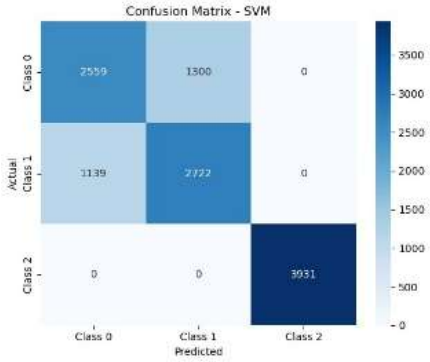

Figure 4. Confusion Matrix - SVM

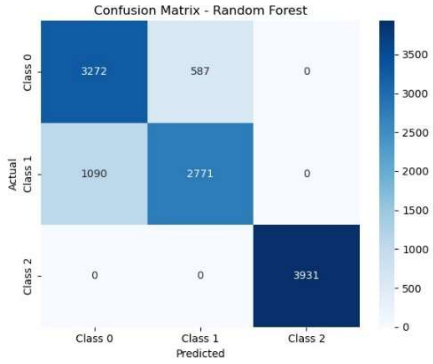

Figure 5. Confusion Matrix – Random Forest

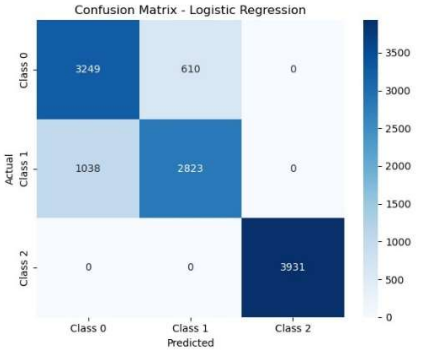

Figure 6. Confusion Matrix – Logistic Regression

VI. CONCLUSION

The objective was to measure and compare the results generated by four machine learning algorithms: XGBoost, Random Forest, Logistic Regression, and SVM, for credit risk classification. XGBoost outcompeted the other algorithms in all the main measures, accuracy (86.27%), precision, recall, F1 score, and R^2 score (0.7946). Logistic Regression and Random Forest performed almost as well as each other, though SVM was not as accurate. Results supported by metrics and confusion matrices indicate that XGBoost performs better than other methods for structured classification problems, like predicting credit risk. This comparison stresses the value of

testing several algorithms to find the one that works best for a specific task. The knowledge gathered here can support future comparisons of machine learning models designed for financial risk analysis.

ACKNOWLEDGEMENT

Y. Jeevan Nagendra Kumar obtained an M.Tech. in Computer Science Technology from Andhra University in 2005 and a Ph.D. in Computer Science and Engineering from Acharya Nagarjuna University in Guntur, AP in 2017. He has been a professor in GRIET's Department of Information Technology since 2005. In addition to his current position as Head of the Department of IT, he has held positions as Dean of the Technology and Innovation Cell, Convener and Vice President of the MHRD Institution Innovation Council, and Coordinator for Robotics, x-Kernel, NAAC, NBA, J-Hub, ARIIA, and the Institute Level CII Survey. He works on web technologies, full stack development, data science, AI & ML, and data mining. He is an Indian Society for Technical Education (ISTE) life member. He has authored numerous articles for reputable international publications and conferences.

REFERENCES

[1] Y. Zhuang and H. Wei, "Design of a personal credit risk prediction model and legal prevention of financial risks," IEEE Access, vol. 12, pp. 146244–146255, 2024.
[2] Z. Hassani, M. A. Meybodi, and V. Hajihashemi, "Credit risk assessment using learning algorithms for feature selection," Fuzzy Inf. Eng., vol. 12, pp. 529–544, 2020.
[3] D. Mienye and N. Jere, "Deep learning for credit card fraud detection: A review of algorithms, challenges, and solutions," IEEE Access, vol. 12, pp. 96893–96910, 2024.
[4] J. C. Alejandrino, J. J. P. Bolacoy, and J. V. B. Murcia, "Supervised and unsupervised data mining approaches in loan default prediction," Int. J. Electr. Comput. Eng., vol. 13, pp. 1837–1847, 2023.
[5] S. B. Coşkun and M. Turanlı, "Credit risk analysis using boosting methods," JAMSI, vol. 19, pp. 5–18, 2023.
[6] K. Wang, M. Li, J. Cheng, X. Zhou, and G. Li, "Research on personal credit risk evaluation based on XGBoost," Procedia Comput. Sci., vol. 199, pp. 1128–1135, 2022.
[7] D. Dansana et al., "Analyzing the impact of loan features on bank loan prediction using Random Forest algorithm," Eng. Rep., vol. 6, e12707, 2024.