

PREDICTING LOAN APPROVAL USING ML ALGORITHMS

¹ M.Priyanka, ² Yerra Sathvika, ³ Challa Saiteja, ⁴ Deevoju Jaya Siri, ⁵ Abhilash

¹Assistant Professor, ²³⁴⁵Students

Department of ECE

Siddhartha Institute of Technology & Sciences, Narapally

priyankamerugu@siddhartha.org.in, 24tq1a0442@siddhartha.co.in, 24tq1a0447@siddhartha.co.in,

24tq1a0449@siddhartha.co.in, 25tq5a0404@siddhartha.org.in

ABSTRACT

Banks and finance companies receive a large number of loan applications every day, and each application must be examined to decide whether the applicant is likely to repay. The decision involves checking income, employment, existing obligations, credit history, the value of any collateral, and the purpose of the loan. When this assessment is performed manually, it takes time, depends on the experience of individual officers, and can produce inconsistent outcomes for applicants with similar profiles. This paper presents a system that predicts loan approval using machine learning algorithms, with the aim of supporting faster and more consistent preliminary decisions.

The system uses publicly available loan application datasets containing applicant details such as gender, marital status, number of dependants, education, employment type, applicant and co-applicant income, loan amount, loan term, credit history, and property area, together with the final approval status. The raw data contains missing values, skewed income distributions, and categorical fields that must be converted into numerical form. A preprocessing stage imputes missing values using suitable strategies for each field, applies log transformation to skewed amounts, and encodes categorical variables. Derived features such as total household income, equated monthly instalment, and the ratio of instalment to income are created because they reflect repayment capacity more directly than the raw values.

Several classification algorithms are trained and compared, namely Logistic Regression, K-Nearest Neighbours, Naive Bayes, Decision Tree, Random Forest, Support Vector Machine, Gradient Boosting, and XGBoost. The imbalance between approved and rejected applications is handled with class weights and oversampling of the minority class. Hyperparameters are tuned using stratified cross-validation, and the models are evaluated on a held-out test set using accuracy, precision, recall, F1-score, and ROC-AUC. The results show that Random Forest and Gradient Boosting achieve the best overall performance, and that credit history, the instalment-to-income ratio, and total income are the most influential features.

The selected model is deployed through a web application built with Flask, where a user enters applicant details and receives a predicted decision along with the probability of approval and the main factors that influenced the result. An administrator view shows summary statistics of processed applications, the distribution of predicted outcomes, and the performance of the model on recent data.

The system is designed to assist loan officers by providing a quick and consistent first assessment, reducing processing time and highlighting applications that need closer review. The final decision remains with the officer. Future work includes fairness checks across applicant groups, the use of richer credit bureau information, and integration with digital document verification.

I. INTRODUCTION

Loans enable individuals to buy homes, vehicles, and education, and they allow small businesses to purchase equipment and manage working capital. For lenders, lending is the main source of income, but it also carries the risk that a borrower will not repay. Every loan application must therefore be assessed carefully to balance the need to serve customers with the need to control losses. As the volume of applications has grown with the expansion of retail and digital lending, the pressure on lenders to make these assessments quickly and accurately has increased.

In many institutions, the initial assessment of an application is still carried out largely by hand. A loan officer reviews the application form and supporting documents, checks the applicant's income and credit history, calculates repayment capacity, and applies eligibility rules. This process can take several days, particularly when applications arrive in large numbers. Decisions may also vary between officers and branches, since each officer applies personal judgement to borderline cases. Applicants who are rejected often receive little explanation, and those who are approved may have waited longer than necessary. For small-value personal and consumer loans, the cost of this manual effort can be large relative to the income the loan generates, which makes it difficult for lenders to serve such customers profitably.

Machine learning can learn from records of past applications and their outcomes to identify the combinations of applicant characteristics associated with approval or rejection. Classification algorithms such as logistic regression, decision trees, random forests, and gradient boosting can process a new application in a fraction of a second and produce a consistent prediction with an associated probability. When combined with methods for explaining individual predictions, these models can also indicate which factors most affected the outcome, supporting transparency for both the lender and the applicant.

This paper describes a loan approval prediction system that prepares application data through careful cleaning and feature engineering, trains and compares a range of classification algorithms, and deploys the best model in a web application. The objective is to provide lenders with a reliable preliminary assessment that speeds up processing, improves consistency, and directs officer attention to cases where human judgement is most needed, while keeping the final decision in human hands. The work also shows how a complete machine learning workflow, from exploratory analysis to a deployed web service, can be organised in Python using widely available open-source libraries.

II. LITERATURE SURVEY

The use of statistical methods to support lending decisions dates back several decades, when lenders began replacing purely judgemental assessment with numerical scoring. Early credit scoring models used linear discriminant analysis to separate good and bad applicants based on characteristics such as income, age, occupation, and time at current address. Logistic regression later became the preferred technique because it provides a direct estimate of the probability of a favourable outcome and does not require the strict distributional assumptions of discriminant analysis. Studies in this period demonstrated that statistical scorecards reduced default rates and processing time compared with manual assessment, and that they produced more consistent decisions across branches. Researchers also noted that such models depend heavily on the quality and representativeness of historical data, and that a model trained on past approvals may not reflect the behaviour of applicants who were rejected and therefore never observed as borrowers.

With the growth of machine learning, many studies compared a wider range of classifiers for loan approval and credit assessment. Decision trees were valued for producing rules that can be read and understood, while k-nearest neighbours and naive Bayes were used as simple and fast baselines. Support vector machines were reported to perform well on moderate-sized datasets with appropriate kernel selection. Comparative studies on public loan datasets frequently found that ensemble methods, particularly random forests and gradient boosting, achieved the highest accuracy and ROC-AUC, because they combine many weak learners and are less sensitive to noise and outliers than single trees. At the same time, several authors reported that the difference between well-tuned logistic regression and ensemble methods was small on some datasets, suggesting that data preparation and feature design can matter as much as the choice of algorithm.

Data preprocessing has been identified as a key factor in the performance of loan prediction models. Loan application datasets commonly contain missing values in fields such as credit history, loan amount, and self-employment status, and studies compared imputation methods ranging from simple mean and mode substitution to model-based approaches. Income and loan amount distributions are typically highly skewed, and log transformation was found to improve the performance of models that are sensitive to scale, such as logistic regression and support vector machines. Feature engineering based on financial reasoning, such as combining applicant and co-applicant income, computing the equated monthly instalment, and calculating the ratio of instalment to income, was repeatedly shown to improve prediction, since these features capture repayment capacity more directly than individual raw fields.

Class imbalance and evaluation methodology have received considerable attention. In many loan datasets, approved applications outnumber rejected ones, and a classifier that simply predicts approval for everyone can achieve deceptively high accuracy. Studies applied oversampling methods such as SMOTE, undersampling of the majority class, and cost-sensitive learning to address this, and recommended evaluation using precision, recall, F1-score, and ROC-AUC rather than accuracy alone. The use of stratified cross-validation to obtain stable performance estimates and to tune hyperparameters without touching the test set was widely advised. Some researchers also highlighted the unequal cost of different errors, since approving an

applicant who later defaults usually costs the lender more than rejecting an applicant who would have repaid, and suggested adjusting decision thresholds accordingly.

Recent research has focused on explainability, fairness, and deployment. Regulations in many countries require lenders to provide reasons for adverse decisions, and methods such as feature importance, partial dependence plots, and SHAP values have been used to explain the predictions of complex models. Studies on fairness have examined whether loan models treat applicants differently based on gender, marital status, or location, and have proposed measures and techniques to detect and reduce such bias. Work on deployment has described web and mobile applications that allow officers or applicants to obtain instant preliminary decisions, with emphasis on data validation, security, and monitoring of model performance over time. The present system draws on these directions by combining careful preprocessing, financially meaningful features, comparison of multiple classifiers, imbalance handling, explanation of predictions, and a web interface for practical use.

III. SYSTEM ANALYSIS

The system is analysed as a sequence of steps that converts an application record into a predicted decision. Historical application data, containing applicant demographics, income, loan details, credit history, property area, and the final approval status, is loaded and examined through exploratory analysis to understand distributions, missing values, and relationships between variables and the outcome. The preprocessing stage imputes missing categorical values with the most frequent category and missing numerical values with the median, applies log transformation to income and loan amount, and encodes categorical fields using one-hot or label encoding. Derived features including total income, equated monthly instalment, instalment-to-income ratio, and loan amount per dependant are computed. The dataset is split into training and test sets with stratification, and class imbalance in the training set is handled through class weights and oversampling. Eight classification algorithms are trained and tuned with cross-validation, and the best model is selected by F1-score and ROC-AUC. In deployment, the web application validates user input, applies the same preprocessing pipeline, and returns the prediction, probability, and key contributing factors.

Existing System

In the existing process followed by many lenders, a loan application is submitted on paper or through an online form and then passed to a loan officer for manual verification. The officer checks identity and income documents, obtains a credit report, calculates repayment capacity using standard ratios, and compares the application against the lender's eligibility rules. Borderline cases are escalated to senior officers, and the decision is communicated to the applicant after several days or weeks. Where automated scoring is used, it is often a simple rule-based system that applies fixed cut-offs to individual variables, such as minimum income or minimum credit score, without considering how different factors interact. Such rules can reject applicants who are strong in other respects or approve applicants whose combination of factors indicates risk. Decisions vary between officers and branches, and the reasons for

rejection are rarely explained in detail. Historical application data is stored but is seldom analysed to improve future decisions.

Disadvantages of Existing System

- Manual verification and assessment take several days for each application.
- Decisions vary between officers and branches for similar applicants.
- Rule-based systems apply fixed cut-offs to individual variables in isolation.
- Interactions between income, loan amount, and credit history are not considered.
- Large volumes of applications create backlogs during busy periods.
- Applicants receive little explanation of the reasons for rejection.
- Historical application data is not used to improve future decisions.
- Borderline cases require repeated escalation to senior officers.
- Human fatigue and bias can affect the quality of assessments.
- Processing cost per application remains high for small-value loans.

Proposed System

The proposed system predicts loan approval using machine learning algorithms and presents the result through a web application. Historical application records are cleaned by imputing missing values, transforming skewed income and loan amounts, and encoding categorical fields. Features that reflect repayment capacity, including total income, equated monthly instalment, instalment-to-income ratio, and loan amount per dependant, are derived. Logistic Regression, K-Nearest Neighbours, Naive Bayes, Decision Tree, Random Forest, Support Vector Machine, Gradient Boosting, and XGBoost classifiers are trained with class weights and oversampling to address imbalance, and their hyperparameters are tuned through stratified cross-validation. Models are compared using accuracy, precision, recall, F1-score, and ROC-AUC on a held-out test set, and the best model is saved together with its preprocessing pipeline. The Flask web application accepts applicant details, validates them, and returns a predicted decision with its probability and the main contributing factors computed with SHAP values. An administrator panel summarises processed applications and tracks model performance.

Advantages of Proposed System

- Provides a preliminary loan decision within seconds of data entry.
- Produces consistent predictions for applicants with similar profiles.
- Considers interactions between income, loan amount, and credit history.
- Uses financially meaningful features such as instalment-to-income ratio.
- Compares eight classification algorithms and selects the best performer.
- Handles class imbalance to avoid bias towards the majority outcome.
- Explains each prediction by showing its most influential factors.
- Reduces workload and processing time for loan officers.
- Keeps a record of predictions for monitoring and later retraining.
- Keeps the final decision with the loan officer for accountability.

IV. METHODOLOGY

The methodology begins with the collection of a public loan application dataset and exploratory data analysis using summary statistics, histograms, box plots, and cross-tabulations of each variable against the approval status. Missing values are identified and imputed, with mode imputation for categorical fields and median imputation for numerical fields, and credit history is treated carefully since it has a strong effect on the outcome. Income and loan amount are log-transformed to reduce skewness, and categorical variables are encoded. Derived features are then computed, with the equated monthly instalment obtained from the loan amount, term, and an assumed interest rate. The data is split into training and test sets in an eighty to twenty ratio with stratification. SMOTE is applied to the training data only, and numerical features are standardised for distance-based and margin-based models. Each classifier is tuned with grid search and five-fold stratified cross-validation. The tuned models are evaluated on the test set using accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices, and the chosen model is serialised and loaded by the web application.

SYSTEM ARCHITECTURE

The architecture of the system is organised into layers that carry data from application records to the user interface. The Data Source Layer holds historical application records and receives new applications entered through the web form. The Data Preprocessing Layer imputes missing values, transforms skewed variables, and encodes categories. The Feature Engineering Layer derives total income, instalment, instalment-to-income ratio, and related features. The Imbalance Handling Layer applies class weights and oversampling to the training data. The Model Training Layer trains the eight classifiers and tunes their hyperparameters. The Evaluation and Selection Layer compares the models and chooses the best one. The Prediction Service Layer loads the saved model and pipeline and scores new applications. The Explanation Layer computes the main factors behind each prediction, and the Web Interface Layer presents the input form, prediction result, explanation, and administrator statistics.

Predicting loan approval using ML algorithms

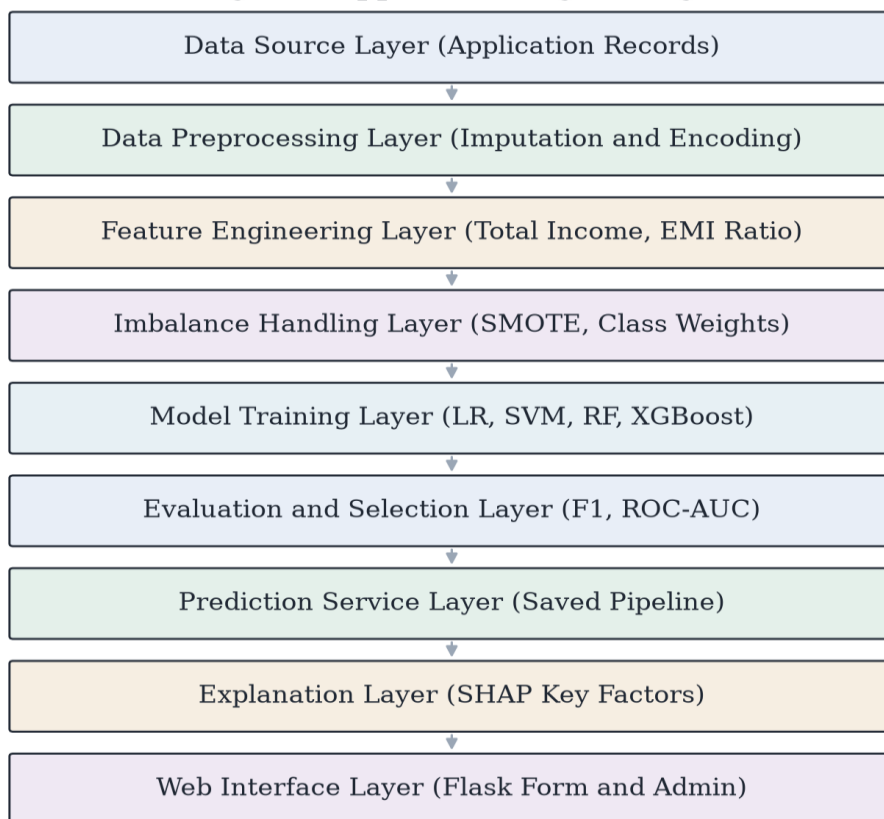
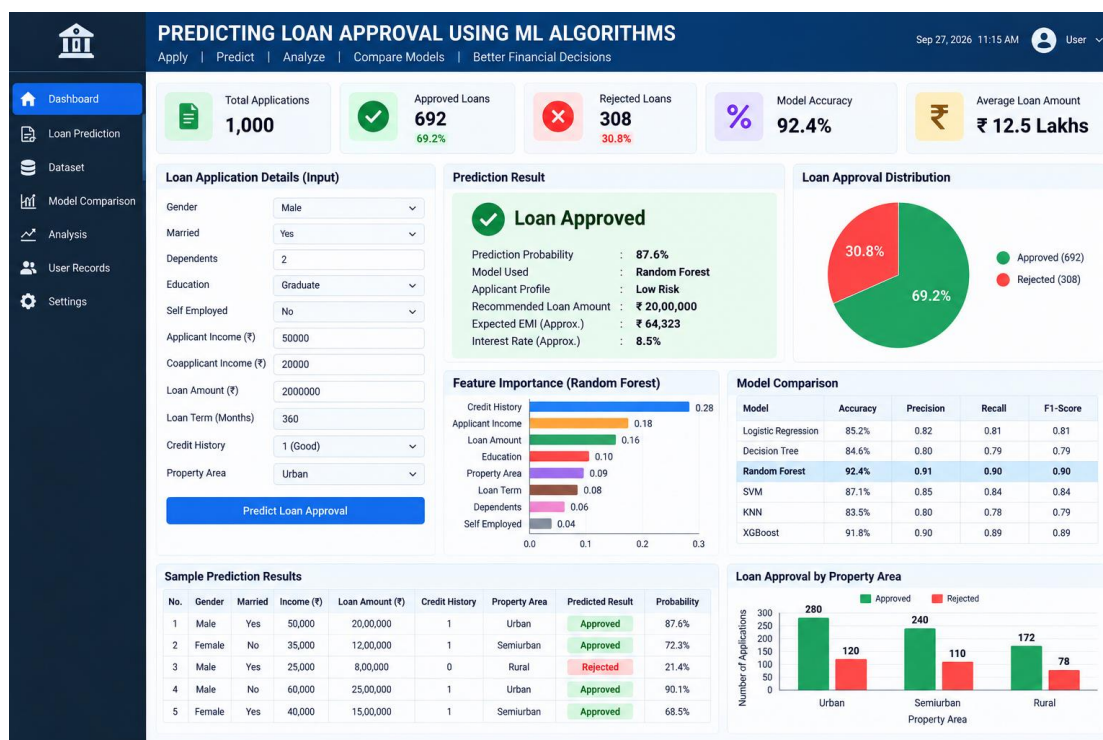


Fig. 1 System Architecture

V. RESULT AND OUTPUT



VI. CONCLUSION

This paper has presented a loan approval prediction system that uses machine learning algorithms to provide a fast and consistent preliminary assessment of loan applications. By learning from records of past applications and their outcomes, the system identifies the combinations of applicant characteristics associated with approval and applies this knowledge to new applications within seconds.

Data preparation and feature engineering played a central role in achieving good performance. Imputing missing values carefully, transforming skewed income and loan amounts, and deriving features such as total income and the instalment-to-income ratio gave the models a clearer picture of repayment capacity. Among the eight algorithms compared, Random Forest and Gradient Boosting achieved the best balance of precision and recall, and credit history emerged as the single most influential factor, followed by repayment capacity measures.

The web application makes the model practical to use. Loan officers can enter applicant details and immediately obtain a predicted decision, a probability, and the main factors behind the prediction. This explanation helps officers understand and verify the model's reasoning, identify borderline cases that deserve closer review, and communicate reasons to applicants. The final decision remains with the officer, so the system supports human judgement rather than replacing it. Recording every prediction together with the eventual decision also creates a useful record for auditing the model and for identifying cases where officers and the model disagree.

Future work can include systematic fairness assessment to ensure that applicants are not treated differently on the basis of gender, marital status, or location, the use of richer credit bureau and banking transaction data, and integration with digital document verification to reduce manual checks further. Regular monitoring and retraining would help the model remain accurate as applicant profiles and economic conditions change over time.

REFERENCES

- [1] Kumar, R. D., Prudhviraaj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In *Handbook of Artificial Intelligence and Wearables* (pp. 145-158). CRC Press.
- [2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In *The International Conference on Artificial Intelligence and Smart Environment* (pp. 557-564). Cham: Springer Nature Switzerland.
- [3] Sv satykrishna, shirisha rangu, bhargavi nalacheruve. (2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm
- [4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.

-
- [5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time Object Detection in Drone Surveillance Using YOLOv5," in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.
- [6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, "Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks," in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.
- [7] R. D. Kumar, V. N. S. Manaswini, "Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology," in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.
- [8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie recommender using collaborative filtering and sentiment analysis," International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.
- [9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system using LightGBM and decision tree machine learning models," Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.
- [10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, "Smart agriculture through IoT and machine learning for analyzing carbon footprints," in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.
- [11] Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer learning approach: MobileNetV2 with CNN," SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.