ADHIL PAYINGALIL
12/16/24

Loan Approval Prediction – A Comparative Study Using Machine Learning Algorithms

I. ABSTRACT

This study addresses the challenge of loan approval prediction using machine learning techniques, focusing on the impact of preprocessing and feature selection. The analysis utilizes the "Loan Status Prediction" dataset, which includes applicant attributes such as income, credit history, and loan terms. Comprehensive preprocessing steps, including feature encoding, scaling, and oversampling with SMOTE, were employed to address data inconsistencies and class imbalance. Recursive Feature Elimination (RFE) was applied to identify nine key predictors, reducing dimensionality. Five machine learning models—Support Vector Machine (SVM), Naive Bayes, Random Forest, Logistic Regression, and Decision Tree—were trained and evaluated using metrics such as accuracy, precision, recall, and F1-score.

Results demonstrate that models trained on all features generally outperformed those trained on RFE-selected features. SVM, for instance, achieved an accuracy of 74.19% and an F1-score of 71.08% with all features, compared to 69.89% accuracy and 67.34% F1-score with RFE. Similarly, Naive Bayes showed higher accuracy and F1-scores when using all features. However, Logistic Regression exhibited improved performance with RFE, achieving 73.12% accuracy and a 70.14% F1-score compared to 70.97% accuracy and a 68.27% F1-score with all features. These findings highlight the nuanced trade-offs of feature selection, emphasizing that its impact can vary across different algorithms. Overall, the study underscores the importance of comprehensive preprocessing and tailored feature selection in developing effective predictive models for financial decision-making.

Keywords—Machine Learning, Feature Selection, Data Processing, Loan Default prediction, Banking, Commercial Lending

II. Introduction

Loan approval prediction is a critical task in the financial industry, where institutions must evaluate the risks that are associated with approving or rejecting loan applications. The decision-making process often involves analyzing multiple applicant attributes, such as income, credit history, and loan terms, which significantly influence the likelihood of repayment. With the increasing availability of structured data, machine learning offers a powerful approach to automating this evaluation process, enabling faster and more accurate decisions.

The objective of this study is to develop a robust predictive model for loan approval using a real-world dataset sourced from Kaggle. The dataset comprises 615 records and 13 features capturing various applicant characteristics. Through a systematic application of data preprocessing techniques, feature engineering, and machine learning algorithms, this project aims to address key challenges such as class imbalance and feature irrelevance.

This research is particularly relevant to the financial industry, where effective loan approval systems can enhance operational efficiency, reduce default rates, and improve customer satisfaction. By leveraging machine learning models, financial institutions can move toward more data-driven and objective decision-making processes.

The latter part of this report discusses the methodology, including data preprocessing, model training, and evaluation. A comparative analysis of five machine learning algorithms highlights their performance, emphasizing the importance of feature selection and preprocessing in improving predictive accuracy.

.

III. LITERATURE REVIEW

Many people seek bank loans, but banks have limited resources and can only extend credit to a select number of customers. The credit provided can either become an asset for the bank, generating income through interest, or a liability if the borrower fails to repay the loan. A significant amount of disbursed capital can turn into bad debt if the bank lacks adequate information about the borrower's repayment ability. Therefore, predicting which customers are likely to repay loans is a concept of extreme importance for banks and other institutions in financial service industry to decrease risk and increase profitability (1). Moreover, the correct loan approval operations increase customer loyalty, and satisfaction, and expands the customer base (5).

There are numerous studies that uses various models to predict loan default by customers. For example, In the study conducted by Saini (2), The explored algorithms include Random Forest Classifier, K-Nearest Neighbors Classifier, Support Vector Classifier, and Logistic Regression. According to the results of the study, the RF algorithm was the most successful algorithm with an accuracy rate of 98.04%. With the increasing demand for loans, banks are forced to lend despite their limited resources. Another study explored the use of artificial intelligence models to predict the safety of loan applications by mining data from banks past lending experiences. This approach aims to enhance the safe lending process by conserving banks' efforts and resources. The research utilized the SVM algorithm and achieved an accuracy rate of 81%. (6). The study by Ramachandra et al. aims to deploy the model on cloud-based platforms using machine learning algorithms and concepts to understand and identify the working methods of loan systems for prediction. The main objective of the project is to predict which customers will or will not repay their loans, using leading algorithms such as Decision Tree (DT), Logistic Regression (LR), and Random Forest (RF). The LR algorithm achieved an accuracy of 86% with minimal error.

Two critical issues are associated with the use of machine learning in the presence of huge and noisy data, namely **dimensionality curse** and **class imbalance** (3). To address the dimensionality curse, several studies employ feature selection methods. A study by Sinap, 2024 used K-Best, and Recursive Feature elimination methods to handle the class imbalance. The models generated by Recursive Feature elimination showed up with more accuracy, furthermore, the same study used SMOTE to address the class imbalance. The models in the study achieved more than 90% accuracy.

While many research papers are comparative, evaluating multiple algorithms using metrics like accuracy, few have discussed feature selection and class imbalances. Our research will employ methods like Recursive Feature Elimination (RFE) for feature selection and compare five machine learning algorithms with and without feature elimination to determine the best one.

IV. METHODS

This section outlines the preprocessing steps, feature selection techniques, and machine learning models implemented for loan approval prediction

A. Feature Encoding

To prepare the data for machine learning models, categorical features were transformed into numerical representations using the following techniques:

1) Label Encoding: Used for features with a natural ordinal relationship, such as Gender, Married, Education, and Self_Employed. Example transformations include:

Gender: Female → 0, Male → 1

Married: No → 0, Yes → 1

2) One-Hot Encoding: Applied to features without inherent ordinal relationships, such as Dependents, which was expanded into binary columns (e.g., Dependents_0, Dependents_1, etc.).

3) Derived Features: A binary feature Long_Term_Loan, was created to represent loans exceeding 360 months (→ 1) and others (→ 0).

B. Feature Scaling

Numerical features (ApplicantIncome, CoapplicantIncome, and LoanAmount) were standardized to ensure uniform contribution during model training. Standardization adjusted the mean to 0 and the variance to 1, reducing the influence of larger magnitude features and improving performance for distance-based algorithms like Support Vector Machines (SVM).

C. Resampling – SMOTE

Class imbalance in the target variable (Loan_Status) was addressed using the **Synthetic Minority Oversampling Technique (SMOTE)**. By generating synthetic samples for the minority class (loan rejections), SMOTE ensured a balanced class distribution, enhancing model performance for minority class predictions. Importantly, SMOTE was applied only to the training dataset after splitting to avoid data leakage.

D. Feature Selection – RFE

To identify the most informative features, **Recursive Feature Elimination (RFE)** was employed using Logistic Regression as the base estimator. RFE iteratively removed the least important features, ultimately selecting nine critical predictors:
Married, Self_Employed, Credit_History, Property_Area, Dependents_0, Dependents_1, Dependents_2, Dependents_3, Long_Term_Loan
These features were used in subsequent model training and evaluation.

E. Classification Models

Five classification algorithms were implemented to predict loan approval outcomes:

- **Logistic Regression:** A linear model for binary classification.
- **Random Forest:** An ensemble method using decision trees.
- **Support Vector Machine (SVM):** A model that finds the hyperplane maximizing class separation.
- **Decision Tree:** A non-parametric tree-based model.
- **Naive Bayes:** A probabilistic model based on Bayes' theorem.

Each model was trained using the full set of preprocessed features and the reduced feature set selected via RFE, enabling a comparative analysis of feature selection's impact.

F. Performance Metrics

The models were evaluated using the following metrics:

- **Accuracy:** Overall correctness of predictions.
- **Precision:** Proportion of correctly predicted positive cases.
- **Recall:** Proportion of actual positive cases correctly predicted.
- **F1-Score:** Harmonic mean of precision and recall.
- **Area Under the Curve (AUC):** Measures the ability to distinguish between classes.

1) Confusion Matrix: Confusion matrices were analyzed to understand the breakdown of:
True Positives (TP)
True Negatives (TN)
False Positives (FP)
False Negatives (FN)
This analysis highlighted the models' strengths and weaknesses in predicting each class.

2) ROC Curve: Receiver Operating Characteristic (ROC) curves and AUC values were used to assess the models' discrimination ability between positive and negative classes. Models achieving higher AUC values demonstrated better classification performance.

This project utilizes the "Loan Status Prediction" dataset, sourced from a Kaggle, to develop machine learning models capable of predicting loan approval outcomes. This dataset is representative of real-world loan application data used by financial institutions. It includes various applicant characteristics such as income, loan amount, credit history, and co-applicant information, which are key factors in the loan approval process. The dataset consists of 615 records, each described by 13 features. The features and their corresponding descriptions are detailed in **Table 1.**

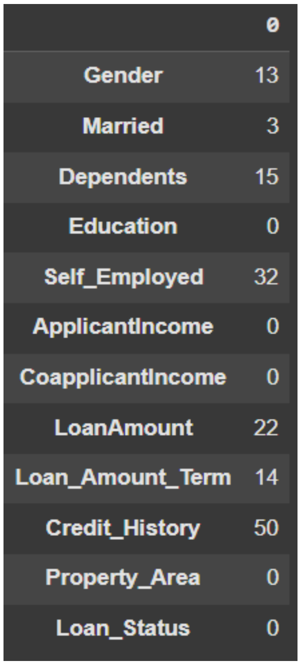
TABLE I. DATASET FEATURES AND TYPES

Variable	Description	Data Type
Loan_ID	Unique Loan ID (to be dropped)	Categorical
Gender	Applicant's gender (Male/Female)	Categorical
Married	Applicant's marital status (Yes/No)	Categorical
Dependents	Number of dependents	Categorical (Ordinal)
Education	Applicant's education level (Graduate/Undergraduate)	Categorical
Self_Employed	Applicant's employment status (Self-employed/Not self-employed)	Categorical
ApplicantIncome	Applicant's income	Numerical
CoapplicantIncome	Co-applicant's income	Numerical

LoanAmount	Loan amount in thousands	Numerical
Loan_Amount_Term	Term of the loan in months	Numerical
Credit_History	Credit history meets guidelines (0 or 1)	Numerical (Binary)
Property_Area	Applicant's property area (Urban/Semi Urban/Rural)	Categorical
Loan_Status	Loan approval status (Yes/No) - Target Variable	Categorical (Binary)

V. DATA PREPERATION

Initial exploratory data analysis (EDA) was conducted to gain a comprehensive understanding of the dataset's characteristics and identify potential issues or patterns. This analysis revealed several key observations. Missing values were found to be present in several features, including Gender, Married, Dependents, Self Employed, Credit History, Loan Amount, and Loan Amount Term as shown in Table 1. Abbreviations and Acronyms



	0
Gender	13
Married	3
Dependents	15
Education	0
Self_Employed	32
ApplicantIncome	0
CoapplicantIncome	0
LoanAmount	22
Loan_Amount_Term	14
Credit_History	50
Property_Area	0
Loan_Status	0

Fig. 1. Null values by features

The distribution of the Loan Amount Term feature was found to be non-uniform, with a strong concentration of values around 360 months. Furthermore, a scale discrepancy was observed between the Loan Amount feature and the Applicant Income and Co applicant Income features. Finally, the distribution of Loan Amount Term exhibited right-skewness, with most values concentrated at 180, 360, and 480 months. Visualizations, including histograms of Loan Amount Term, Loan Amount, Applicant Income, and Coapplicant Income, along with bar charts of the categorical features and a visualization of missing values, were generated to support these observations.

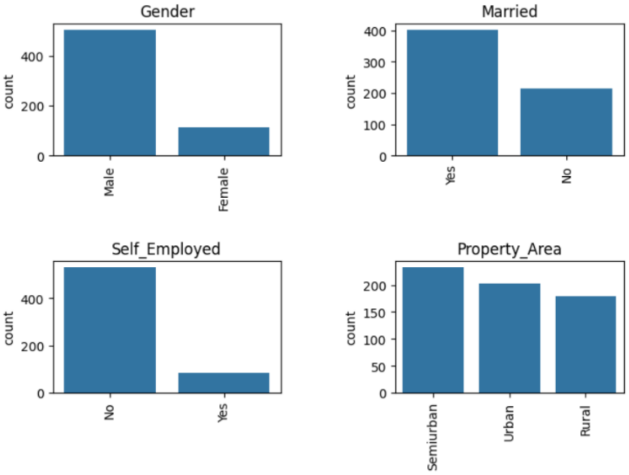


Fig. 2. Categorical variables value counts I

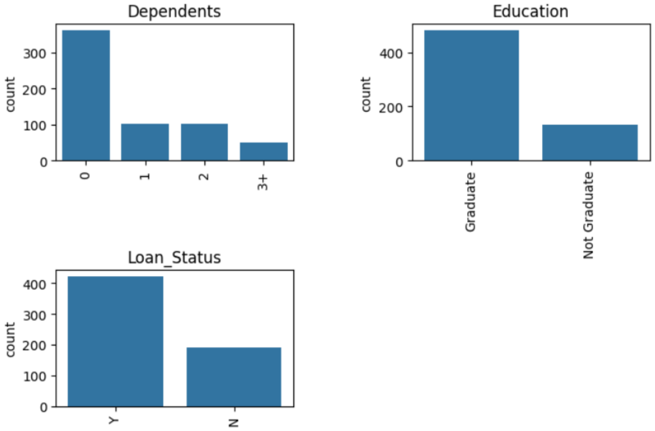


Fig. 3. Categorical variables value counts II

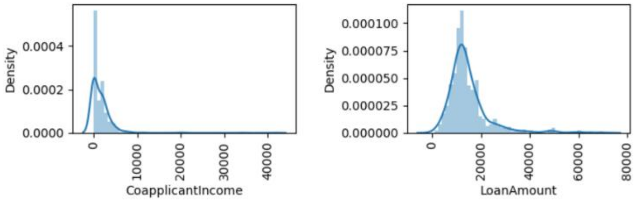


Fig. 4. Numeric values distribution I

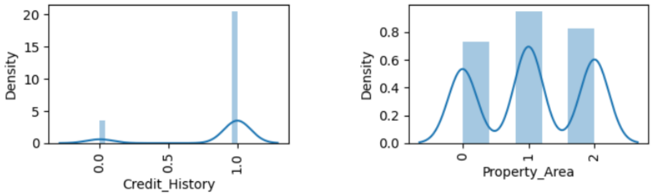


Fig. 5. Numeric values distribution II

A. Target Variable Analysis (Class Imbalance):

EDA also revealed a class imbalance in the target variable, Loan Status. The number of loan approvals significantly outnumbered the number of loan rejections. This imbalance can bias machine learning models towards the majority class, leading to poor performance in predicting the minority class.

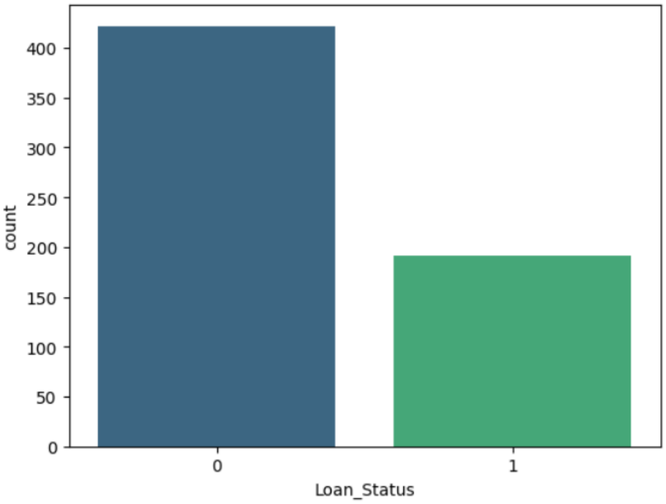


Fig. 6. The target variable visualized – noticeable class imbalance

VI. Data Processing

Based on the insights derived from the EDA, a series of preprocessing steps were implemented to prepare the data for machine learning model training.:

Initially, the Loan ID column, being a unique identifier with no predictive value, was dropped. To address the issue of missing values, mode imputation was employed for the categorical features (Gender, Married, Self Employed, and Credit History), replacing missing entries with the most frequent value for each respective feature. For the numerical features (Loan Amount and Loan Amount Term), median imputation was used, as the median is less sensitive to outliers than the mean.

A. Feature Encoding

Subsequently, feature encoding was performed to convert categorical variables into numerical representations. Label encoding was applied as follows

Gender ('Female' -> 0, 'Male' -> 1)

Married ('No' -> 0, 'Yes' -> 1)

Education ('Graduate' -> 0, 'Not Graduate' -> 1)

Self Employed ('No' -> 0, 'Yes' -> 1)

Property Area ('Rural' -> 0, 'Semiurban' -> 1, 'Urban' -> 2)

Target variable Loan Status ('N' -> 0, 'Y' -> 1).

Loan Amount Term (“1 year or longer” -> 1, ‘shorter than a year’ -> 0)

The Dependents feature, originally represented as 0, 1, 2, and 3+, was first mapped to numerical values 0, 1, 2, and 3, respectively. Then, to avoid imposing an ordinal relationship between these categories, one-hot encoding (creating dummy variables) was applied to the mapped Dependents feature. This resulted in four new binary features, each representing one of the original categories.

B. Feature Scaling

Feature scaling and engineering were performed to optimize the performance of the machine learning models. Addressing the scale discrepancy observed during EDA, the Loan Amount column, which was initially represented in thousands, was multiplied by 1000 to convert it to the actual loan amount. This adjustment ensured consistent units across the features. Subsequently, standardization, a crucial preprocessing step for many machine learning algorithms, was applied to the numerical features (Applicant Income, Co applicant Income, and the adjusted Loan Amount). Standardization transforms the features to have a mean of zero and a standard deviation of one. This is particularly important in this context because the features had varying scales, with Loan Amount potentially having much larger values than income features. Without scaling, algorithms that rely on distance calculations (such as Support Vector Machines or k-Nearest Neighbours) could be disproportionately influenced by features with larger magnitudes. Standardization prevents this by ensuring that all features contribute equally to the model's learning process.

Additionally, a new binary feature, Long Term Loan, was engineered based on Loan Amount Term: loans with a term greater than or equal to 360 months were labelled as long-term loans (1), while others were labelled as short-term loans (0). This feature aimed to capture the potential impact of loan term duration on loan approval.

C. Addressing Class Imbalance

As identified during EDA, the target variable (Loan_Status) exhibited a class imbalance, with a significantly larger number of loan approvals (1) compared to loan rejections (0). This imbalance can lead to biased machine learning models that favor the majority class, resulting in poor predictive performance for the minority class.

To mitigate this potential bias, oversampling was employed using the SMOTE (Synthetic Minority Over-sampling Technique) method. SMOTE generates synthetic samples for the minority class (0) by interpolating between existing minority class instances. This process effectively increases the number of minority class samples, leading to a more balanced class distribution.

Crucially, SMOTE was applied after splitting the data into training and testing sets. This is a critical step to prevent data leakage, where information from the testing set inadvertently influences the training process, leading to overly optimistic performance estimates. The data was split using an 85/15 split, allocating 85% of the data for training and 15% for testing.

After applying SMOTE to the training set, the distribution of Loan Status was balanced, with equal numbers of (0) and (1) instances. This balanced distribution is visualized in **Figure 7**.

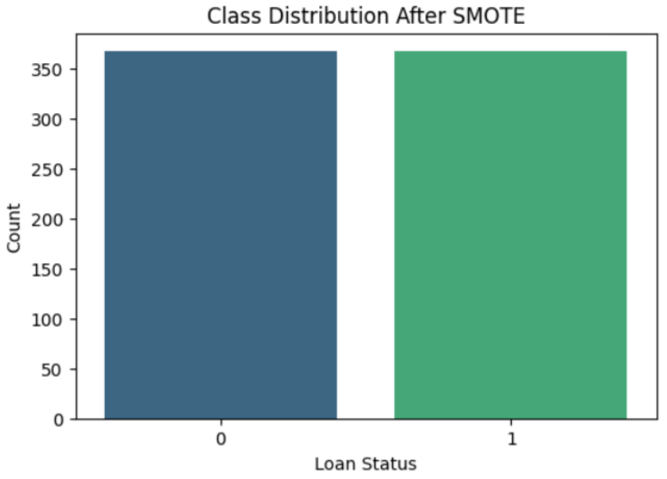


Fig. 7.   Target variable after using SMOTE

VII. Feature Selection

As part of Exploratory data analysis (EDA), the loan approval rates across various applicant characteristics where examined. Analysis of approval rates within each category revealed that certain groups experienced higher approval rates. These included male applicants (81.76%), married applicants (65.31%), graduates (78.18%), those with a positive credit history (85.50%), and those applying for long-term loans (88.11%). Figure 2, and 3 shows the the graphs showing the loan application success of the applicants according to the features.

A correlation analysis further explored the relationships between features and Loan Status. This analysis highlighted a strong positive correlation between Credit History and Loan Status (0.541), confirming the substantial influence of credit history on approval outcomes. Weak positive correlations were also observed with Married (0.091), Dependents_2 (0.062), and Property Area (0.032). These weak correlations are consistent with the observed trends in approval rates within these categories. Other features exhibited negligible correlations, suggesting limited direct linear influence on loan outcomes. The correlation matrix showing all the correlation between the features is shown in Figure 8.

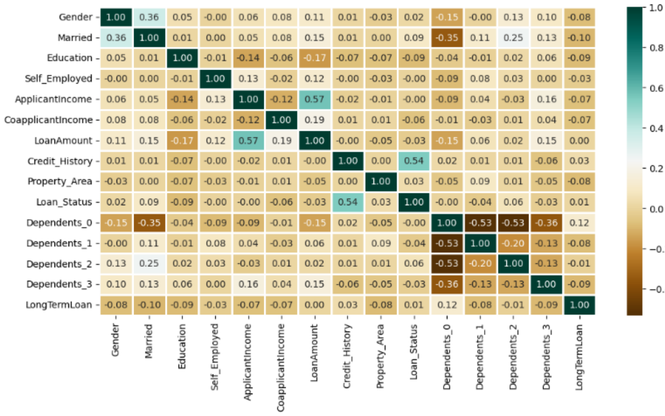


Fig. 8.   Correlation Matrix among features

After looking at EDA, and Correlation we moved on to further feature selection. Feature selection is a crucial process in machine learning, aimed at identifying the most relevant features from a dataset for model training. This process offers several benefits, including preventing overfitting by reducing model complexity, improving model generalization by focusing on informative features, decreasing computational time by eliminating unnecessary features, and enhancing model interpretability.

While correlation analysis can be a useful tool for feature selection, its applicability is limited when dealing with datasets containing a significant proportion of categorical variables, as is the case in this project. Therefore, we focused on analysing the relationship between individual features and the target variable (Loan Status) and employed Recursive Feature Elimination (RFE) as our primary feature selection method.

RFE is a feature selection technique that recursively removes features based on their importance to a chosen model. It works by training the model on the initial set of features and ranking them based on their contribution. The least important feature(s) are then removed, and the process is repeated until the desired number of features is reached. In this project, RFE, using Logistic Regression as the base estimator, was used to select the top features for predicting loan approval.

After running RFE, the following nine features were selected as the most important: Married, Self Employed, Credit History, Property Area, Dependents 0, Dependents 1, Dependents 2, Dependents 3, and Long-Term Loan. These selected features were then used for subsequent model training and evaluation. To maintain consistency and avoid introducing bias, the same train-test split generated during the RFE process was used for all further model development and testing.

VIII. Model Setups

This section outlines the machine learning models employed for loan approval prediction and the experimental setup used for their evaluation. Five classification algorithms were used: Logistic Regression, Random Forest, Support

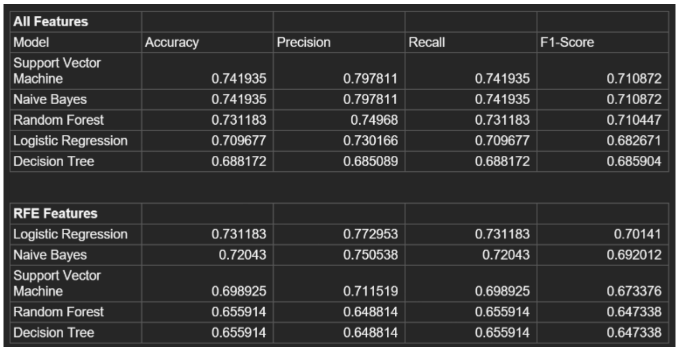
Vector Classifier (SVC), Decision Tree, and Naive Bayes. Detailed descriptions of these models are provided in a later section.

To assess the impact of feature selection, each model was trained and evaluated using two distinct feature sets: the nine features selected by RFE and all pre-processed features. This comparison allowed us to quantify the effect of feature selection on model performance.

All models were trained using the resampled training data generated by the SMOTE oversampling technique to address the class imbalance in the target variable. The same train-test split generated during the RFE process was consistently used across all models and feature sets to ensure a fair and consistent comparison.

IX. Model Comparison and Evaluation

This section presents a comparative analysis of the performance of the five machine learning models (Logistic Regression, Random Forest, SVC, Decision Tree, and Naive Bayes) trained for loan approval prediction. The models were evaluated using accuracy, precision, recall, and F1-score metrics on both the full feature set and the reduced feature set selected by RFE. The results are summarized in Figure 9.



All Features				
Model	Accuracy	Precision	Recall	F1-Score
Support Vector Machine	0.741935	0.797811	0.741935	0.710872
Naive Bayes	0.741935	0.797811	0.741935	0.710872
Random Forest	0.731183	0.74968	0.731183	0.710447
Logistic Regression	0.709677	0.730166	0.709677	0.682671
Decision Tree	0.688172	0.685089	0.688172	0.685904
RFE Features				
Logistic Regression	0.731183	0.772953	0.731183	0.70141
Naive Bayes	0.72043	0.750538	0.72043	0.692012
Support Vector Machine	0.698925	0.711519	0.698925	0.673376
Random Forest	0.655914	0.648814	0.655914	0.647338
Decision Tree	0.655914	0.648814	0.655914	0.647338

Fig. 9. Performance metric by model by feature selection method

Figure 9 summarizes the results which indicate that SVM and Naive Bayes achieved the highest accuracy (0.7419) when using all features. This suggests that these models can effectively utilize all available information, even with potential noise. However, RFE significantly impacted most models. Logistic Regression maintained good performance with the reduced set, while Random Forest and Decision Tree showed a noticeable drop, indicating they might benefit from the removed features. Surprisingly, SVM and Naive Bayes also had decreased accuracy with fewer features, possibly because those features provided useful context .

A. *Comparing the Confusion Matrices:*

Comparing the confusion matrices for models trained with all features versus those using RFE-selected features reveals the nuanced impact of feature selection on loan approval predictions. While using all features generally led to slightly higher overall accuracy, the differences were not substantial. Notably, using fewer features often resulted in a slight increase in false positives (incorrectly predicting loan approval) but a decrease or no change in false negatives (incorrectly predicting loan rejection). This suggests a potential trade-off between overall accuracy and the risk of approving undeserving loans. Individual models showed varying responses to feature selection: SVM, for example, significantly reduced false negatives with RFE, while Naive Bayes saw an increase in false positives.

B. *AUC Analysis*

The Area Under the ROC Curve (AUC) measures a model's ability to distinguish between approved and rejected loan applications. 1 Higher AUC values indicate better discrimination. 2 Our analysis revealed that Random Forest with all features achieved the highest AUC (0.71), suggesting the strongest discriminatory power among the models. However, Random Forest also showed a noticeable decrease in AUC after applying Recursive Feature Elimination (RFE), dropping to 0.63. This indicates that the removed features might be important for Random Forest's ability to effectively separate loan classes. Conversely, the other models (Logistic Regression, Decision Tree, and Naive Bayes) showed relatively consistent AUC values across both feature sets, suggesting their discriminatory power was less affected by feature selection. **Figure 10** shows the ROC curve for both cases with all features and selected features from RFE. **Table 2** shows

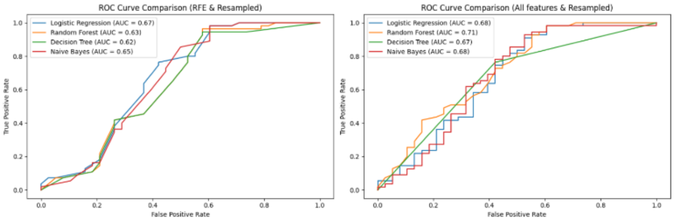


Fig. 10. ROC, and AUC metrics graphs by feature selection methods

TABLE II. AUC Values

AUC Values		
Model	RFE Features	All Features
Logistic Regression	0.67	0.68
Random Forest	0.63	0.71
Decision Tree	0.62	0.67
Naive Bayes	0.65	0.68

X. Results & Discussion

The comparative analysis of five machine learning models—Logistic Regression, Random Forest, Support Vector Machine (SVM), Decision Tree, and Naive Bayes—evaluated their effectiveness in predicting loan approval outcomes. These models were assessed using accuracy, precision, recall, F1-score, and Area Under the Curve (AUC) metrics on both the full feature set and the reduced feature set obtained through Recursive Feature Elimination (RFE). The results highlight how feature selection influenced the performance of different algorithms.

A. Performance Summary

Table X summarizes the models' performance metrics with all features and the RFE-selected feature set. Key findings include:

- **SVM and Naive Bayes:** Achieved the highest accuracy (0.7419) using all features, effectively leveraging the entire feature set despite potential noise.
- **Logistic Regression:** Maintained strong performance across both feature sets, with only a slight reduction in metrics after feature selection, demonstrating its robustness to feature reduction.
- **Random Forest and Decision Tree:** Displayed a significant drop in performance with RFE features, indicating reliance on the removed features for optimal classification.

B. Impact of Feature Selection

- **SVM:** While achieving high accuracy with all features, SVM experienced a performance decline with RFE-selected features (accuracy dropped to 0.6989). This indicates that some excluded features contributed to its performance, even amidst potential noise. However, RFE did reduce false negatives, suggesting improved recall for loan approval predictions.
- **Naive Bayes:** Although Naive Bayes performed well with all features, feature selection resulted in a slight reduction in accuracy (from 0.7419 to 0.7204) and an increase in false positives, highlighting its reliance on a broader feature set to maintain balance.
- **Logistic Regression:** Demonstrated consistent performance, with F1-scores of 0.7157 (all features) and 0.7014 (RFE features). Its ability to retain performance with fewer features underscores its adaptability to changes in feature dimensionality.
- **Random Forest and Decision Tree:** Both models experienced noticeable declines in accuracy and AUC with the reduced feature set. Random Forest's AUC dropped from 0.71 to 0.63, indicating that the removed features were critical for distinguishing between loan classes. This highlights the sensitivity of tree-based models to feature availability.

The results underscore the varying sensitivity of machine learning models to feature selection. Tree-based models like Random Forest and Decision Tree suffered notable declines with fewer features, emphasizing their dependence on comprehensive feature availability. In contrast, Logistic Regression and Naive Bayes demonstrated resilience, retaining robust performance with the reduced feature set. SVM exhibited the most significant trade-off, as feature selection improved its recall and reduced false negatives but came at the cost of reduced overall accuracy.

These findings suggest that feature selection, while valuable in simplifying models and reducing computational complexity, must be tailored to the specific model characteristics. For applications like loan approval prediction, balancing the trade-offs between model complexity, interpretability, and predictive accuracy is crucial.

XI. Limitations of the Study

While this study successfully explores loan approval prediction using machine learning, several limitations highlight areas for future improvement. One significant shortcoming is the lack of hyperparameter tuning, such as grid search or random search, which could have significantly improved model performance. For example, models like Support Vector Machines (SVM) and Random Forest are highly sensitive to hyperparameter choices, and tuning parameters like the kernel type in SVM or the number of trees in Random Forest could have led to better results. Without this optimization, the model performance presented here might not reflect their full potential.

Another limitation stems from the use of Recursive Feature Elimination (RFE) with Logistic Regression as the base model for feature selection. While Logistic Regression is a robust and interpretable choice, its linear nature might bias the feature selection process, potentially overlooking non-linear interactions important for models like Random Forest or SVM. This could explain why some models, such as Random Forest, performed worse when using RFE-selected features compared to the full feature set. Future research could explore model-agnostic feature selection methods, such as SHAP (Shapley Additive exPlanations) or permutation importance, which provide a more comprehensive analysis of feature contributions.

Additionally, the study relied on a relatively small dataset with only 615 records. While this dataset is representative of real-world applications, its size limits the generalizability of the findings, especially for machine learning models that often benefit from larger training samples. Small datasets can exacerbate issues like overfitting, particularly in models with high complexity, such as Random Forest or SVM. Cross-validation was not explicitly detailed in the study, which could have ensured more reliable performance evaluations across different data splits.

The study also primarily focused on a single approach to address class imbalance—SMOTE (Synthetic Minority Oversampling Technique). While effective, SMOTE can introduce synthetic samples that might not fully represent real-world data distributions. Alternative methods, such as ensemble techniques like Balanced Random Forest or undersampling the majority class, could offer additional insights into balancing the dataset while maintaining its integrity.

XII. Conclusion

This study highlights the potential of machine learning in automating and improving loan approval predictions by systematically addressing data challenges like missing values, class imbalance, and feature irrelevance. The project demonstrated that preprocessing techniques, including feature scaling, encoding, and oversampling, are pivotal to model success. Among the tested algorithms, Support Vector Machines (SVM) and Logistic Regression emerged as the most reliable models. SVM excelled in leveraging the full feature set but experienced trade-offs after feature selection, while Logistic Regression showcased consistent robustness across both datasets. These findings underscore the importance of tailoring data preprocessing and feature selection to the strengths of specific algorithms.

Despite its limitations, the study provides a solid foundation for the development of more efficient, accurate, and interpretable predictive models. By streamlining loan approval workflows, such models can empower financial institutions to make faster, more objective, and customer-centric decisions. Future work can address current shortcomings by incorporating advanced optimization techniques, alternative feature selection methods, and interpretability tools to better align machine learning solutions with industry needs.

References

[1] E. Kadam, A. Gupta, S. Jagtap, I. Dubey, and G. Tawde, "Loan approval prediction system using logistic regression and CIBIL score," in 2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC), Jul. 2023, pp. 1317-1321

[2] P. S. Saini, A. Bhatnagar, and L. Rani, "Loan approval prediction using machine learning: A comparative analysis of classification algorithms," in 2023 3rd Int. Conf. Adv. Comput. Innov. Technol. Eng. (ICACITE), May 2023, pp. 1821-1826.

[3] Suliman Mohamed Fati, "A Loan Default Prediction Model Using Machine Learning and Feature Engineering," ICIC Express Letters, ICIC International, vol. 18, no. 1, pp. 27-37, January 2024.

[4] Vahid Sinap, "A Comparative Study of Loan Approval Prediction Using Machine Learning Methods," Gazi University Journal of Science, Part C: Design and Technology, vol. 12, no. 2, pp. 644-663, April 2024.

[5] V. Leninkumar, "The relationship between customer satisfaction and customer trust on customer loyalty," Int. J. Acad. Res. Bus. Soc. Sci., vol. 7, no. 4, pp. 450-465, 2017.

[6] Y. Diwate, P. Rana, and P. Chavan, "Loan Approval Prediction Using Machine Learning," Int. Res. J. Eng. Technol. (IRJET), vol. 8, no. 05, 2021.