Available online at https://icsejournal.com/index.php/JCSE
Journal of Computer Science an Engineering (JCSE)
e-ISSN 2721-0251
Vol. 5, No. 2, August 2024, pp. 64-79

Building Reliable Loan Approval Systems: Leveraging Feature Engineering and Machine Learning

Maryam Shoaeinaeini[1*], Milad Shoaeinaeini[2], Brent Harrison[1], Milad Jasemi[3]

[1]Department of Computer Science, University of Kentucky, Lexington, KY 40506 United States of America
[2] Department of Management and Accounting, Islamic Azad University, Islamshahr Branch, Tehran, 1584743311 Iran
[3] Stephens College of Business, University of Montevallo, AL 35115 United States of America
[1]maryam.shoaei@uky.edu *; [2] milad.shoaei20@gmail.com; [3] brent.harrison@uky.edu, [4] mjasemiz@montevallo.edu
* corresponding author

ARTICLE INFO

Article History:
Received April 09, 2024
Revised May 20, 2024
Accepted June 02, 2024

Keywords:
Loan Approval Prediction,
Feature Engineering
Machine Learning
Ensemble Learning

Correspondence:
E-mail: maryam.shoaei@uky.edu

ABSTRACT

Automating loan approval system is essential in today's banking system. Even with the shift to online platforms, the traditional method is still cumbersome and needs a lot of customer-related data. This study proposes a robust solution to overcome these challenges. Despite previous studies, new financial indicators in feature engineering stage are introduces to extract more important client information, thereby improving prediction robustness and accuracy. To implement our integrated approach, an online dataset from a finance company, is utilized. The dataset is preprocessed by various data preparation techniques, including cleaning, transformation, and feature engineering. Subsequently, the preprocessed data undergoes a range of powerful machine learning techniques such as K-Nearest Neighbor, Decision Tree, Gaussian Naive Bayes, and Logistic Regression. Additionally, three robust ensemble methods including Random Forest, AdaBoost Classifier, and Gradient Boosting Classifier are employed for further improveness in performance. The presented approach succeeded to acheive the highest accuracy by AdaBoost Classifier at 88%. A comparison with the original preprocessed model using ROC curve and feature importance analysis demonstrates the superior performance of our approach, with a larger area under the ROC curve and reduced false positive rate. Furthermore, the comparison findings show a stronger reliance of our model on financial features rather than personal customer features, highlighting its robust classification performance. These results indicate the potential strength of our model to replace the current loan approval system in real-world applications.

1. Introduction

The financial companies is suffering from the challenge lies in accurately evaluating the risk associated with loan applications. Inaccurate obtained risk can result in financial consequences that may jeopardize its long-term sustainability in the current competitive market. Traditionally, these decisions have relied on credit scoring and risk analysis tools; however, there are several cases each year in which borrowers fail to repay the loans and default, causing financial institutions to suffer significant losses [1], [2]. To address this issue, we require a robust model that incorporates essential customer features to automate and expedite the process, enhancing its efficiency and accuracy. Extensive research and numerous studies have been conducted to identify the pivotal variables that impact the prompt repayment of loans. Two distinct categories of customer features play a significant role in loan default scenarios. Firstly, demographic variables form a crucial set of factors that influence default behavior [3]. Secondly, individual characteristics, such as the borrower's attitude, also contribute to the borrower's default risk propensity [4]. Apart from retrieving effective customer features, selecting the right learning model significantly impacts the efficiency, accuracy, and precision of a prediction system. Traditionally, the models that are frequently used for prediction purposes are statistics-oriented models such as Discriminant Analysis (DA) and Logistic Regression (LR) [5], [6]. However, these models are inefficient regarding credit analysis problems because they have limited capacity to handle complex, nonlinear relationships within data [7], [2].

https://doi.org/10.36596/jcse.v5i2.812

Machine learning methods, in contrast, offer solutions to these deficiencies, making them a more advantageous choice for modernizing and improving loan approval systems. Machine learning has recently gained significant traction across various research fields such as traffic [8], energy [9], healthcare [10], and so on. Moreover, machine learning has made significant advancements in financial research, particularly within systems such as loan approval [11], [2]. Moreover, their reliance on insufficient feature engineering and inability to adapt to changing conditions render them impractical and inefficient models. However, the previous machine learning models for the loan approval problem did not hold sufficient attention to the data preprocessing stage, particularly feature engineering, rendering them impractical and inefficient models. In pursuit of automating the loan approval system, this research introduces a robust prediction model encompassing feature analysis, data preparation procedures, and the application of a diverse set of powerful machine learning algorithms, including Decision Tree (DT), Gaussian Naive Bayes (GNB), Logistic Regression, and K-Nearest Neighbor (KNN), on the cleaned and transformed data. To further enhance performance and accuracy, three potential ensemble techniques, namely the Random Forest (RF), AdaBoost Classifier, and Gradient Boosting (GB), are also employed.

The paper makes significant contributions compared to previous research in the field. Unlike earlier studies that focused solely on implementing prediction algorithms, this research places special emphasis on refining dataset representation through feature engineering. In this regard, the study introduces crucial financial indicators derived from existing features, which enrich the data representation. The newly devised features have greater potential to explain the variance in the training data. This enhancement significantly improves accuracy, with the introduction of key metrics such as Equated Monthly Installment (EMI), Total Income, and Balance Income. Moreoverto [7] argued for a heightened attention to the potential consequences of false negatives or (false positives in loan approval problems) to better address the challenges faced by lending institutions. Taking this perspective into account, our paper demonstrates a lower false positive rate and higher area under the receiver operating characteristic curve (ROC curve) compared to the previous models. Additionally, by leveraging both classical and ensemble machine learning algorithms, this paper explores the creation of a more accurate prediction model among a variety of machine learning approaches. The incorporation of ensemble methods and comprehensive preprocessing of data contributes to an impressive accuracy rate of 88%. To the best of our knowledge, none of the previous prediction models have achieved such a combination of accuracy and robustness.

This study begins with a review of prior research in Section 2. Section 3 provides a comprehensive explanation of the data preprocessing techniques and a brief introduction to the machine learning methods used. The implementation details of each machine learning method are presented in Section 4, along with a comparison of the accuracy of the employed classifiers against models from previous studies on loan approval problems. Additionally, this section demonstrates the superiority and robust classification performance of the presented model compared to the baseline model. The results are discussed in Section 5, and the paper concludes with Section 6.

2. Related works

In the competitive realm of finance, credit rating has evolved into a crucial tool, drawing increased attention due to advancements in data science and artificial intelligence. Researchers are now directing their efforts toward predicting loans and evaluating credit risks, responding to the growing demand for loans [12]. Unlike the past reliance on expert judgments, the current trend leans towards automated methods. Researchers are increasingly endorsing the utilization of machine learning algorithms and neural networks (NN) for loan problems [1] and risk assessment [13], marking notable progress and setting the stage for further exploration and analysis. An exploration of the literature reveals a common theme across the loan default problem, loan approval problem, and credit risk assessment – all involve analyzing the risk or possibility of granting loans to applicants based

https://doi.org/10.36596/jcse.v5i2.812

on their financial history [14], [3] (such as credit history and income) and personal characteristics [4], [15] (like location, house ownership).

The risk of defaulting on a loan arises when a borrower lags behind on payments for a specified period. The primary objective of the loan default problem is to classify applicants into binary or multi-graded risk classes regarding loan repayment. Numerous studies, particularly based on the Lending Club dataset from Kaggle, have delved into this domain. The Lending Club dataset utilizes features such as employment details, income, credit history, loan amount, loan purpose, and other relevant financial indicators. The lending Club categorizes loan status into nine classes, ranging from default to fully paid status. [16] classified customers to different risk levels associated with defaulting on loans and they named the problem as loan default prediction. They utilized Random Forest and Decision Tree methods, achieving an 80% accuracy rate after preprocessing the data. [17] addressed class imbalance using the SMOTE method for upsampling the data, attaining a 98% accuracy rate with a Random Forest. [1] claimed that the Random Forest method has higher accuracy compared to logistic regression, decision trees, and support vector machines (SVM). They focused on 850 records of the bank's default payment data, achieving 86.17% accuracy using the Extra Trees Classifier. Eight features related to various debt and income and individual characteristics were defined. The performance of each classifier was analyzed using a variety of performance metrics like accuracy, and ROC curve. Several researchers use the term "risk assessment" with the same objective as that of the loan default prediction problem. In this context, lenders routinely assess the credit risk associated with a borrower's commitment to repaying the mortgage [18]. The research conducted by [19] aimed to classify loan risk using a combination of customer behavior and credit history. They collected their dataset from the banking sector in ARFF format. The authors utilized the j48 algorithm and achieved the highest accuracy (78.38%).

The challenge of predicting loan approvals revolves around the task of identifying eligible applicants based on their financial and personal attributes. Several studies have explored this area using various machine-learning techniques. In the study by [20], the Decision Tree model achieved an impressive 81% accuracy on the Dream Housing Finance dataset, determining whether a loan should be granted to the applicant. Missing values and outliers are overcome by the map function in their preprocessing stage. Reference [2] utilized gradient boosting and logistic regression, reporting testing accuracy of 84% and 82%, respectively. They analyzed the features and their impact on the target using bar charts and box plots. Finally, they stated that gradient boosting offers higher accuracy by improving from prior experiences. Similarly, [21] conducted research using the same dataset and experimented with various classic and ensemble machine learning algorithms, such as Logistic Regression, Decision Tree, Adaboost, and Random Forest. Their findings indicated that Decision Tree outperformed the ensemble and other classic algorithms, achieving an accuracy of 82%.

However, it is worth noting that this paper demonstrates the ROC curve, which suggests that their proposed model exhibits a linear ROC curve. This indicates that the model is still in the learning phase and has not yet stabilized, even at the end of the training process. In a comparative analysis by [22], various machine learning methods including K-Nearest Neighbor, Suppor Vector Machine, Logistic Regression, and Random Forest were evaluated, with Random Forest demonstrating the highest accuracy at 81%. However, this paper predominantly focuses on machine learning algorithms and does not delve into the preprocessing stage in their work. A summary of recent research is presented in Table 1 for better comparison.

It is noteworthy that none of these studies explored feature engineering, a process involving the selection or creation of specific features to enhance data representation. In fact, previous papers solely utilized the existing features of the Dream Housing Finance dataset without engaging in any feature engineering in their proposed models. This oversight may have limited the potential for further accuracy improvements in their results. Furthermore, [7] encouraged the future researchers to pay more attention to the potential consequences of false positives in loan approval problems.

https://doi.org/10.36596/jcse.v5i2.812

Table 1. A Brief Comparison of Recent Studies and the Presented Model

Recent Works	Problem	Dataset	Feature Engineering	Methods	Important Features	Best Model & Accuracy
[12]	Credit scoring	Private	No	NN, DT, and LR selected by OptiML	Personal characteristics	94.7% by LR
[1]	Loan default prediction	Private	No	ExtraTree, RF, CatBoost,Light GB, Extreme GB, GB, DT, AdaBoost,SVM, LR, Naive Bayes, KNN, XGBoost, Ridge classifier	Financial characteristics	85% by RF
[13]	Credit scoring	Private	No	LR, XGBoost	Unknown	91% by XG-boost
[3]	Loan default prediction	Lending Club	Yes	NN	Unknown	93% by NN
[16]	Loan default prediction	Lending Club	No	DT, RF	Unknown	80% by RF
[17]	Loan default prediction	Lending Club	Yes	RF, DT, SVM, LR	Financial characteristics	98% by RF
[19]	Loan default prediction	Private	No	j48, Bayes Net and Naive Bayes	Unknown	78% by J48
[20]	Loan Approval Prediction	Dream Housing Finance	No	DT, SVM, KNN, GB	Unknown	81% by DT
[2]	Loan Approval Prediction	Dream Housing Finance	No	LR, GB	Unknown	83% by LR
[21]	Loan Approval Prediction	Dream Housing Finance	No	LR, DT, RF, AdaBoost	Unknown	82% by DT
[22]	Loan Approval Prediction	Dream Housing Finance	No	RF, KNN, SVM, LR	Unknown	81% by RF
Our Work	Loan Approval Prediction	Dream Housing Finance	Yes	DT,KNN,LR, GNB, RF,GB, AdaBoost	Financial characteristics	88% by AdaBoost

In fact, false positives can have serious consequences for lending institutions. When a loan is approved for an applicant mistakenly, it increases the risk of default and may impose financial burden on the lender. Therefore, reducing false positive rate requires higher consideration in loan approval processes for maintaining the financial sustainability. Reviewing these studies reveals significant gaps, particularly the lack of attention to the feature engineering stage, which plays a pivotal role in making robust classification. To address these gaps, this paper proposes new financial indicators as features in loan approval problem, thereby improving prediction robustness. Additionally, the study places emphasis on a comprehensive introduces new financial factors aimed at extracting more relevant customer features, thereby improving prediction robustness. Additionally, the study places emphasis on a comprehensive data cleaning process and feature analysis, which includes outlier detection and assessment of each feature's impact on the target variable. Notably, our study achieves an accuracy of 88% in loan approval prediction by leveraging ensemble machine learning methods. Furthermore, the performance of each algorithm is rigorously evaluated using six metrics. Of particular importance are the feature importance and ROC curve analyses, which highlight the significance of the newly devised features in each method by comparing classification outcomes with and without the inclusion of these features.

https://doi.org/10.36596/jcse.v5i2.812

3. Method

In this section, we elaborate on our methodology, which involves several essential stages. We commence with an exhaustive data collection process to acquire the requisite information. Following this, we delve into a comprehensive feature analysis, categorizing features and uncovering their underlying patterns. Our data preprocessing phase is delineated, encompassing meticulous data cleaning, feature transformation, and feature engineering to refine data representation before pulling data to prediction models. Lastly,we provide various advanced machine learning techniques applied to the preprocessed data, ensuring precise analysis and prediction.

3.1 Data Collection

Customer features and loan status information are extracted from the Dream Housing Finance company dataset available on the Kaggle website [23]. Kaggle, renowned as a prominent online platform, is widely recognized for hosting an extensive and diverse collection of datasets, making it a valuable resource for data-driven projects and analyses. The studied dataset is composed of eight categorical features and three numerical features. The training dataset is utilized to train the model, and it is further partitioned into two segments at an 80:20 ratio. The majority of the dataset is employed for model training, while the smaller portion functions as a test set. As a result, the accuracy of our developed model is evaluated through this testing phase.

3.2 Feature Analysis

Bivariate and multivariate analyses of features are conducted to categorize features and uncover special relationships and patterns, thereby enhancing model performance, accuracy, and precision. These analyses are used in outlier identification and selecting appropriate feature intervals. In this study, categorical features of the dataset are Dependents, Education, Gender, Self_Employed, Married, Credit_History, Property_Area, and Loan_Amount_Term.

A comparative analysis for each categorical feature is conducted based on the Loan Status (target value). As depicted in Figure 1, the majority of loan applicants are male, married, and not self-employed, and they possess a credit history background. Furthermore, this analysis yields valuable information, indicating that graduate applicants with zero dependents in semiurban areas have a high likelihood of loan approval. Additionally, it is noteworthy that the requested loans are typically repaid within 360 days. Last but not least, applicants with a credit history score of zero face significantly reduced chances of loan approval.

Among existing features, applicant income, coapplicant income, and loan amount are categorized as numerical features based on their data type. To illustrate the distribution of data, a box plot demonstration compares numerical features with Loan Status in Figure 2. Besides, a comprehensive statistical analysis is performed, and the information is shown in Table 2.

To retrieve the correlation between features and the positive target value (when the loan is approved), a correlation heatmap is applied. As demonstrated in Figure 3, each feature is assigned to one row and one column, and the intersection values between columns and rows show the correlation among features. Also, each cell has a color such that the strength of color determines the strength of correlation. Apart from the correlation of features, this graph can be highly beneficial in determining the importance of each feature on the target value. For example, It is retrieved that credit_history has the highest effect on the loan approval. This underscores the importance of considering credit history as a pivotal factor in the feature importance analysis of machine learning classifiers, as elaborated further in subsequent sections.

https://doi.org/10.36596/jcse.v5i2.812

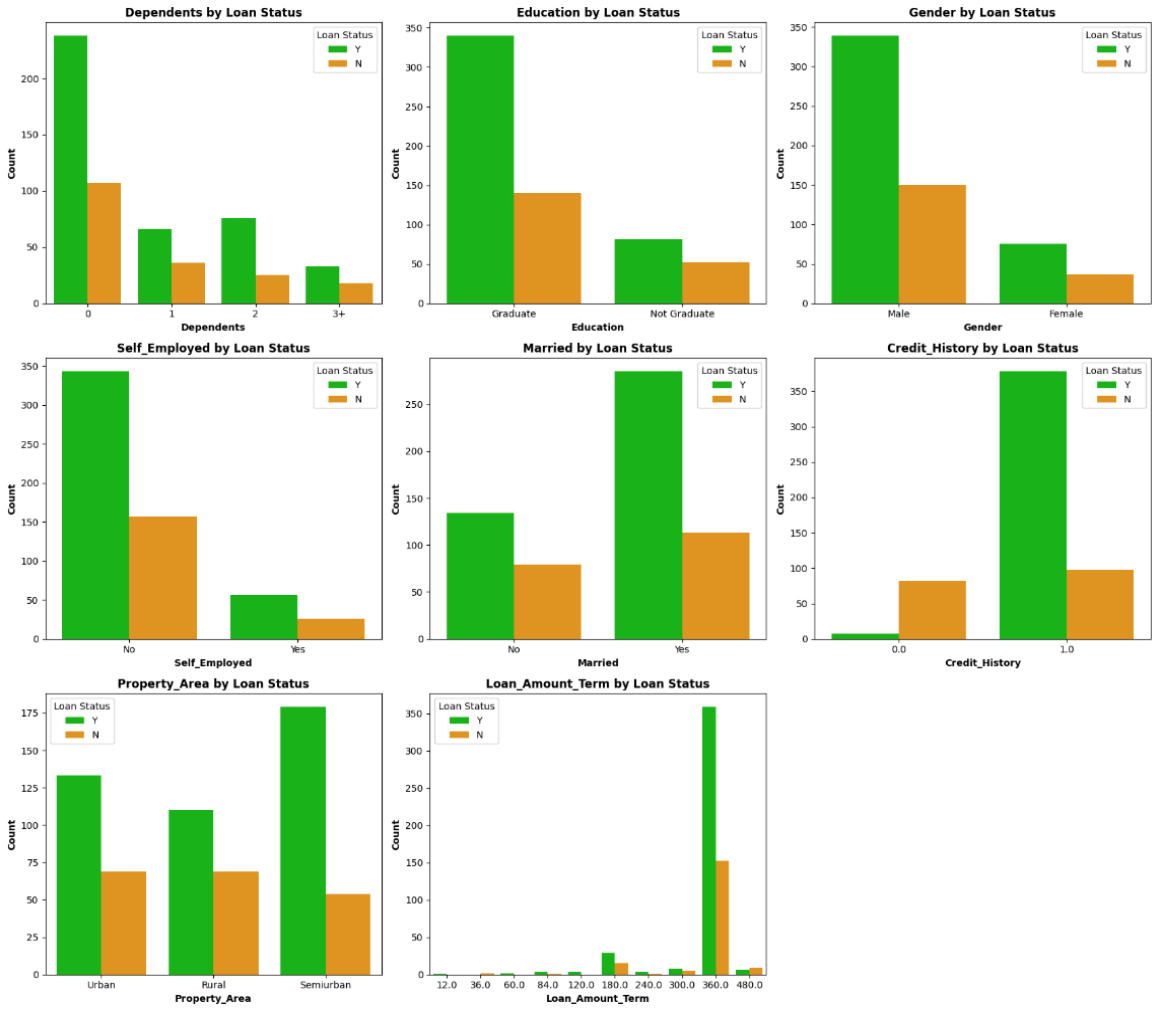

Figure 1. Comparative Analysis of each Categorical Feature versus Loan Status

Tabel 2. Statistical Information of Numerical Features

Statistics	Coapplicant Income	Loan Amount	Applicant Income
count	614.00	592.00	614.00
mean	1621.25	146.41	5403.45
standard deviation	2926.25	85.59	6109.04
minimum	0	9.00	150.00
maximum	41667.00	700.00	81000.00

3.3 Data Preparation

Following the collection and analysis of data, a crucial step involves the preprocessing of data before feeding it into machine learning models as input. This process significantly contributes to achieving higher accuracy in model outcomes. Data preparation encompasses three essential stages: data cleaning, data transformation, and feature engineering. Each stage plays a pivotal role in enhancing the quality and relevance of the data, ensuring optimal performance and effectiveness when utilized in machine learning algorithms.

https://doi.org/10.36596/jcse.v5i2.812

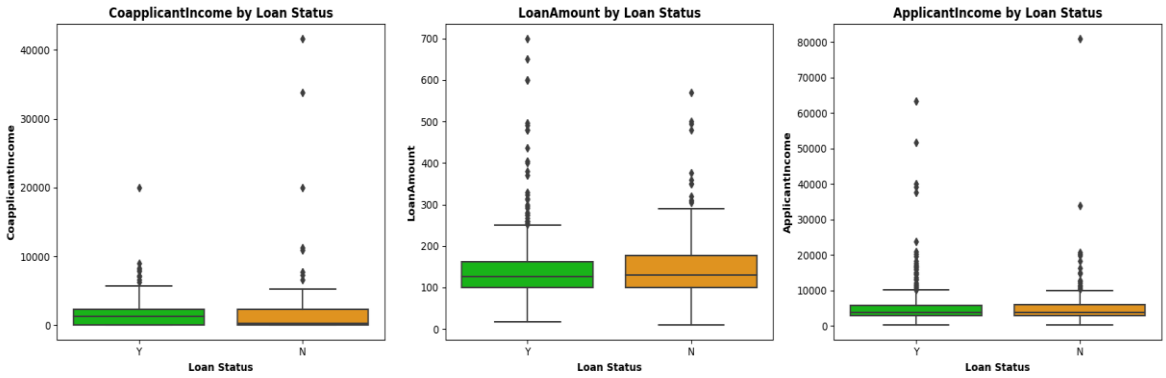

Figure 2. Comparative Analysis of each Numerical Feature versus Loan Status

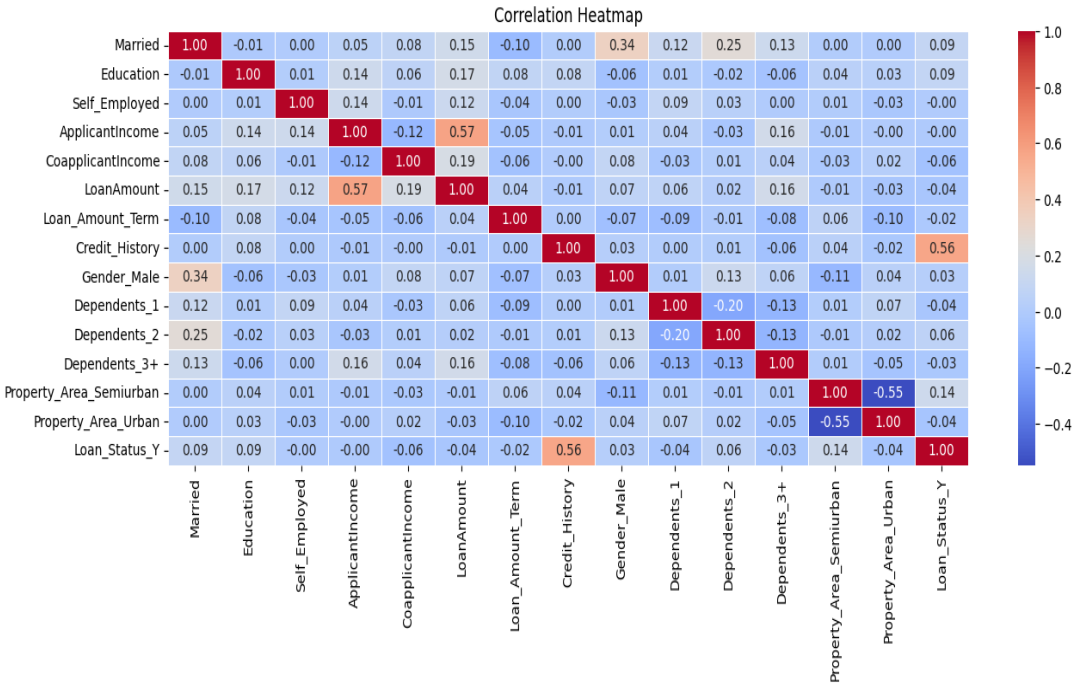

Figure 3. The Demonstration of Correlation Heatmap between Features and the Target Variable

3.3.1 Data Cleaning

In the data cleaning stage, we employ techniques tailored to address missing values based on the nature of the features. Categorical features with missing values are imputed using the most frequently occurring category, ensuring robust data integrity. For numerical feature columns, we utilize the mean value technique to effectively fill in any null or missing values. This meticulous approach to data cleaning enhances the quality and reliability of our dataset, preparing it for subsequent analysis. Our dataset contained null values within categorical features such as ‘Gender’, ‘Married’, ‘Dependents’, ‘Loan_Amount_Term’, ‘Self_Employed’, and ‘Credit_History’. These gaps in the data were addressed by imputing them with the most frequently occurring category for each respective feature. For instance, as the majority of applicants in the dataset were male, missing values in the ’Gender’ column were replaced with ’male.’ Conversely, for numerical features such as ’LoanAmount’, we employed the mean value imputation method, resulting in an average loan amount of $146.41.

3.3.2 Data Transformation

As the second stage of data preparation, a crucial step involves the transformation of categorical features. To achieve this, we employ an approach that distinguishes between nonordinal and ordinal features. For nonordinal features, we implement one-hot encoding, a technique that effectively

https://doi.org/10.36596/jcse.v5i2.812

converts categorical data into a binary format, enabling the model to interpret these variables more comprehensively. Conversely, for ordinal features, we opt for numerical classification, a method that assigns numerical values to the categories based on their inherent order or significance. This meticulous transformation process equips our dataset with the appropriate data format to facilitate accurate and meaningful analysis, setting the stage for robust machine learning model development. In our analyzed dataset, 'Dependents' and 'Property_Area' stand out among the categorical features as ordinal variables. To represent these features, we have created corresponding converted variables: 'Dependents_1', 'Dependents_2', and 'Dependents_3+' for 'Dependents,' and 'Property_Area_Semiurban' and 'Property_Area_Urban' for 'Property_Area'.

3.3.3 Feature Engineering

Given that the profitability of bank loans hinges on customers' capacity to repay the loan in full and on time [24], [25], the loan approval problem is heavily reliant on the financial standing of customers. Thus, it becomes imperative to extract additional financial insights from the features available in the dataset. features primarily capture customer behavior, there is a clear necessity to augment this information with more detailed financial indicators. This enhancement will provide a more comprehensive understanding of the financial standing of customers, thereby facilitating more accurate and informed decisions in the loan approval process.

In the feature engineering phase, we devise pivotal financial indicators that significantly contribute to our dataset's depth and utility. Within this context, we have created essential metrics, including Equated Monthly Installment (EMI), Total Income, and Balance Income, leveraging existing features as the foundation. These newly engineered variables provide valuable insights and enhance the overall robustness of our dataset. EMI is a monthly payment that a borrower makes to a lender at a designated time. It is calculated based on (1).

$$EMI = \frac{LoanAmount}{Loan\_Amount\_Term} \tag{1}$$

In addition, the total income of each applicant is calculated as the sum of the applicant's income and the coapplicant's income, as demonstrated in (2). Therefore, we can introduce a new feature called 'Total_Income' and subsequently remove the redundant features of applicant's income and coapplicant's income based on (2).

$$Total\ Income = Applicant\ Income + CoApplicant\ Income \tag{2}$$

Lastly, we calculate the net income for each applicant by subtracting their monthly EMI payment from their total income, as stated in (3).

$$Net\ Income = Total\ Income - EMI \tag{3}$$

3.4 Machine Learning Techniques

We employ a diverse set of powerful classification methods, encompassing Decision Trees, Gaussian Naive Bayes, Logistic Regression, and K-Nearest neighbor. These algorithms are systematically applied to the data after thorough cleansing and transformation. We also incorporate three resilient ensemble methods to improve performance and accuracy: the Random Forest, AdaBoost Classifier, and GradientBoosting Classifier. Subsequently, we evaluate the performance of each classifier by six computing essential metrics. Finally, We compare our model to a baseline which is based on the features of the original dataset without any feature engineering.

4. Results and Discussion

After meticulously preprocessing our datasets, we executed a comprehensive suite of four classification algorithms, ranging from Decision Trees and Gaussian Naive Bayes to Logistic Regression and K-Nearest Neighbors, all implemented in the Python programming language. Beyond these standalone models, we harnessed the strength of three resilient ensemble methods—

https://doi.org/10.36596/jcse.v5i2.812

namely, the Random Forest, AdaBoost Classifier, and Gradient Boosting Classifier to bolster both performance and accuracy. Our primary focus is on presenting the outcomes of the top-performing five algorithms in this particular scenario. These classifiers were trained on preprocessed datasets. We employed a diverse set of six metrics for a robust evaluation, including Accuracy, F1-Score, Recall, Precision, ROC Area, and Feature Importance. The rationale behind selecting these evaluation metrics is their aptness for classification algorithms, especially when dealing with binary predicted variables. Besides, the ROC curve and Feature Importance were chosen to compare the performance of the proposed prediction model with the model comprising existing features, emphasizing the significance of newly devised features in enhancing classification robustness and accuracy. In this section, a brief explanation of each machine learning method is provided. Following that, the performance metrics are discussed in detail to compare methods.

4.1 Machine Leatning Models

To automate the loan approval process, various advanced machine learning algorithms such as Decision Tree, Gaussian Naive Bayes, Logistic Regression, and K-Nearest Neighbor are implemented on the processed data. Additionally, to improve performance and accuracy, three ensemble methods—Random Forest, AdaBoost Classifier, and Gradient Boosting Classifier—are utilized.

4.1.1 Decision Trees Classifier

Decision trees (DTs) serve as a non-parametric method in supervised learning, useful for both classification and regression tasks. They work by constructing a model based on straightforward decision rules generated from the features of the dataset, aiming to predict the value of a target variable accurately. In this problem, we set the min_samples_leaf to 50 and the max_depth to 3 to ensure comprehensive exploration of the data while maintaining computational feasibility.

4.1.2 Gaussian Naive Bayes

The Gaussian Naive Bayes classifier is a probabilistic machine learning algorithm rooted in Bayes' theorem, operating under the assumption that the features are distributed according to a Gaussian (normal) distribution. This classifier is implemented in Python using the GaussianNB class from the sklearn library.

4.1.3 Logistic Regression

Logistic Regression is a widely used machine learning algorithm for binary classification problems. It models the probability that a given input belongs to a particular class. This method is utilized using LogisticRegression class from sklearn library.

4.1.4 K-Nearest Neighbors

The K-Nearest Neighbors (KNN) classifier is a simple, versatile algorithm for classification and regression. It predicts the class of a data point based on the majority class among its K closest neighbors in the training set, using a distance metric like Euclidean distance. In this study, the number of neighbors is set to 5, following the default setting on the sklearn website.

4.1.5 Random Forest

The Random Forest classifier is an ensemble machine learning method. It combines the predictions of multiple decision trees to produce a more accurate and robust model than any individual tree. This ensemble approach helps to reduce overfitting and improve generalization to new data. Similar to Decision Tree classifier, the max_depth is set to 3, and the number of estimators (n_estimators) is set to 50.

4.1.6 AdaBoost Classifier

The AdaBoost classifier is an ensemble learning algorithm that enhances the performance of weak learners by iteratively focusing on misclassified instances. It assigns higher weights to these instances in subsequent iterations to boost their importance. Its key parameter is the number of weak learners

https://doi.org/10.36596/jcse.v5i2.812

(n_estimators), determining the number of iterations the algorithm performs. AdaBoost is advantageous for its effectiveness in handling high-dimensional data and its ability to mitigate overfitting. The n_estimators parameter in the Random Forest classifier and the n_estimators parameter in the AdaBoost classifier serve a similar purpose—they both control the number of base estimators (weak learners) in the ensemble. Thus, the n_estimators parameter in the AdaBoost classifier is set to 50, following its default value on the scikit-learn website [26].

4.1.7 Gradient Boosting Classifier

The Gradient Boosting classifier is a popular ensemble learning method that builds a strong predictive model by combining multiple weak learners, typically decision trees, in a sequential manner. Gradient Boosting fits each new model to the residual errors made by the previous models, thereby gradually reducing the errors in predictions. Its key parameters include the number of trees (n_estimators), controlling the number of boosting stages. Like AdaBoost classifier, the n_estimator is set to 50 in gradient boosting classifier.

4.2 Performance Metrics

After implementing various machine learning classifiers, performance metrics are essential for evaluating the effectiveness and generalization ability of each classifier. In this section, performance metrics including accuracy, precision, recall, and F1 score are presented to indicate the effectiveness of each model. Additionally, the presented model is compared with the baseline model using feature importance and the ROC curve to highlight the superiority and robustness of our model in the loan approval problem.

4.2.1 Accuracy

Accuracy gauges the proportion of correctly predicted categorical values, representing the percentage of accurate predictions in a given classification model. The formula to calculate the accuracy is as below:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Where TP represents True Positive, TN stands for True Negative, FN corresponds to False Negative, and FP indicates False Positive. The accuracy of each method is represented in Table 3. Referring to this Table, AdaBoost stands out as the most accurate classifier with an 88% accuracy rate, while K-nearest neighbor exhibits the lowest accuracy among the classifiers, at 68%. Additionally, Figure 4 illustrates the accuracy of each method, facilitating a comparative analysis.

4.2.2 Precision

Precision denotes the ratio of accurately predicted positive classes to the total projected positive classes, providing insight into the model's ability to precisely identify positive instances. In fact, It quantifies the precision of positive predictions within the model's output. The precision metric for each method is represented in Table 3. The formula to calculate the precision is as follows:

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

https://doi.org/10.36596/jcse.v5i2.812

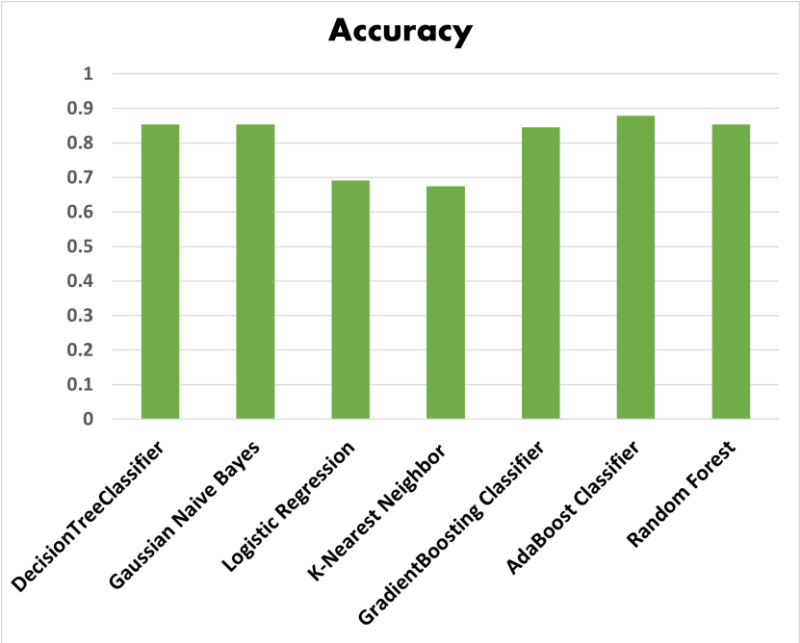

Figure 4. An Evaluation of the Accuracy Metric Obtained from the Machine Learning Methods

Tabel 3. Performance Metrics for the Studied Methods

Methods	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.85	0.83	0.99	0.90
GaussianNB	0.85	0.84	0.98	0.90
Logestic Regression	0.69	0.69	0.98	0.90
K-Nearest Neighbour	0.68	0.71	0.89	0.79
Gradient Boosting	0.85	0.84	0.97	0.90
AdaBoost	0.88	0.87	0.97	0.92
Random Forest	0.85	0.84	0.98	0.90

4.2.3 Recall

Recall, often referred to as True Positive Rate, measures the proportion of correctly predicted positive values among all actual positive instances. It quantifies the model's effectiveness in capturing all positive outcomes. The recall metric for the considered method is represented in Table 3. The recall metric is formulated as below:

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

4.2.4 F1 Score

The F1 score reflects the balanced performance of a classification model by considering both precision and recall simultaneously. This metric is calculated for each method as shown in Table 3. It is calculated as below:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{7}$$

4.2.5 ROC Curve

In the context of classification problems, the Receiver Operating Characteristic (ROC) curve is a graphical representation that illustrates the trade-off between the true positive rate (sensitivity) and false positive rate (1-specificity) at various thresholds. [7] argued for a need for greater attention to the potential consequences of false negatives or (false positives in loan approval problems) to better address the challenges faced by lending institutions. The ROC curve explicitly shows the relationship

https://doi.org/10.36596/jcse.v5i2.812

between sensitivity (true positive rate) and the false positive rate. The Area Under the Curve (AUC) serves as a summary measure, with a higher AUC indicating better overall model performance in distinguishing between classes. Our baseline model, which utilizes the features of the existing dataset without undergoing any feature engineering, serves as our point of reference for comparison. To comparison of these two models, the top three most accurate models, namely Random Forest, Gradient Boosting, and AdaBoost, have been selected and illustrated in Figures 5 to 7. In the Figures, the performance of both models is compared to the naive model which does not employ any sophisticated algorithms to make predictions and often resorts to random or constant predictions. It serves as a initial model for comparison, highlighting the performance improvements achieved by more advanced classification methods.

As the ROC curve advances and the classification threshold is adjusted based on the Figures, it is observed how changes impact the false positive rate. For example, in Figure 7, AdaBoost showcases the robustness of the proposed model compared to the initial model by stabilizing the curve sooner, exhibiting fewer fluctuations, and possessing a larger AUC. Additionally, to facilitate a better comparison of the models, the AUC value has been calculated and is presented in Table 4.

Tabel 4. Area under the ROC curve of the Proposed Model and Baseline Model

Methods	Baseline Model	Proposed Model
Gradient Boosting	0.76	0.82
AdaBoost	0.76	0.80
Random Forest	0.79	0.83

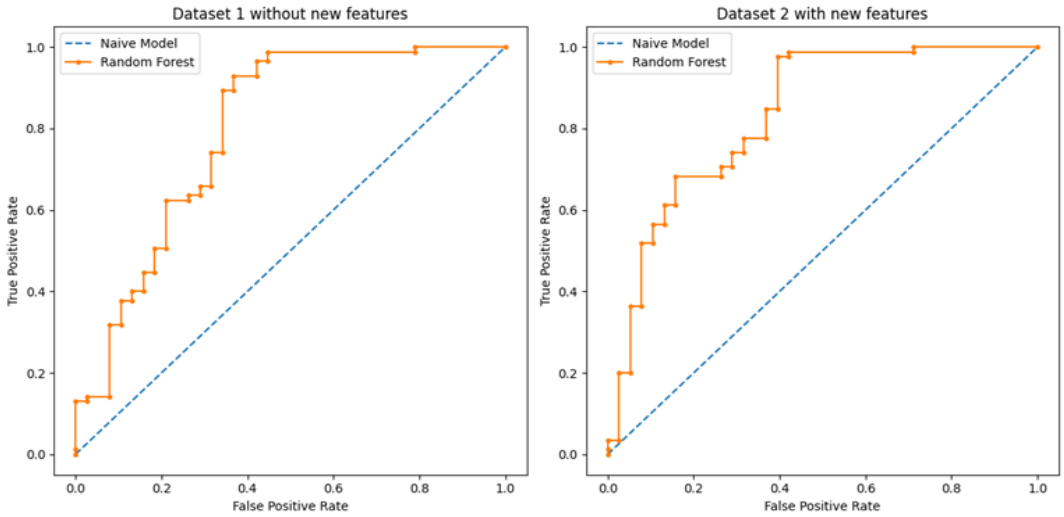

Figure 5. ROC Curves for Random Forest in the Baseline Model and the Proposed Model

4.2.6 Feature Importance

Through this investigation, crucial elements contributing to the accurate prediction of loan approval by the classifiers have been realized, offering valuable insights for decision makers in financial institutions. As depicted in Figures 8 to 10, the principal features influencing the predictions of the top three algorithms in the baseline model and the proposed model are visually presented. The results highlight the significant influence of financial factors like EMI, Balance Income, Total Income, and credit history on the proposed model's outcomes. This emphasizes the critical role played by the newly introduced features derived during the feature engineering phase. However, the Figures reveal that the baseline models exhibit a high dependency on customer behavior features like 'Dependents_3+' and 'Dependents_1', contrary to the findings of the correlation heatmap graph presented in Section 3. This discrepancy suggests that the models may be overfitting and their results may lack robustness and reliability.

https://doi.org/10.36596/jcse.v5i2.812

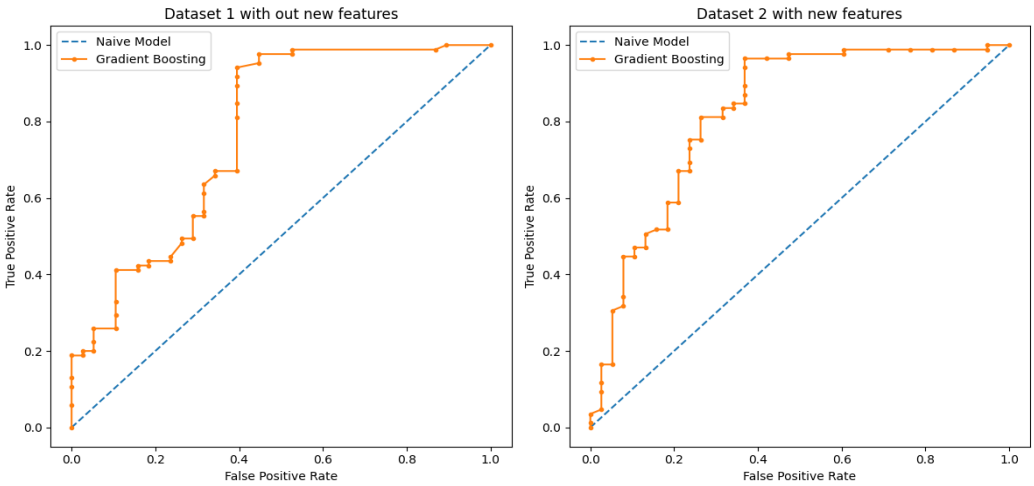

Figure 6. ROC Curves for Gradient Boosting in the Baseline Model and the Proposed Model

5. Discussion

Based on the results obtained in the previous section, it is evident that most classifiers in the proposed model demonstrated significant performance, achieving over 85% accuracy, with the highest accuracy of 88% achieved by the Adaboost classifier. A baseline model, which incorporates the existing dataset's features without any feature engineering, serves as our reference point for comparison. The comparison of the ROC curves associated with each classifier provides valuable insights into their accuracy compared to the baseline model. Given that false positive rates in loan problems can lead to financial losses and reputational damage for lending institutions, it is crucial to analyze these curves. Upon comparing these ROC curves, it becomes evident that the AUC value is greater in the proposed model, consequently resulting in a lower false positive rate compared to the primary model. This indicates that the proposed model offers improved performance in correctly identifying positive cases while minimizing the occurrence of false positives, thus reducing the financial risks associated with lending decisions. Furthermore, the comparison of feature importance of these models provides valuable insights on each model basis. Based on the result, the proposed model demonstrated a high reliance on financial indicators, while the baseline model highly relies on low-correlated customer behavior features with the target, such as 'Dependent_3'. Thus, the classification based on the baseline model is less reliable and accurate.

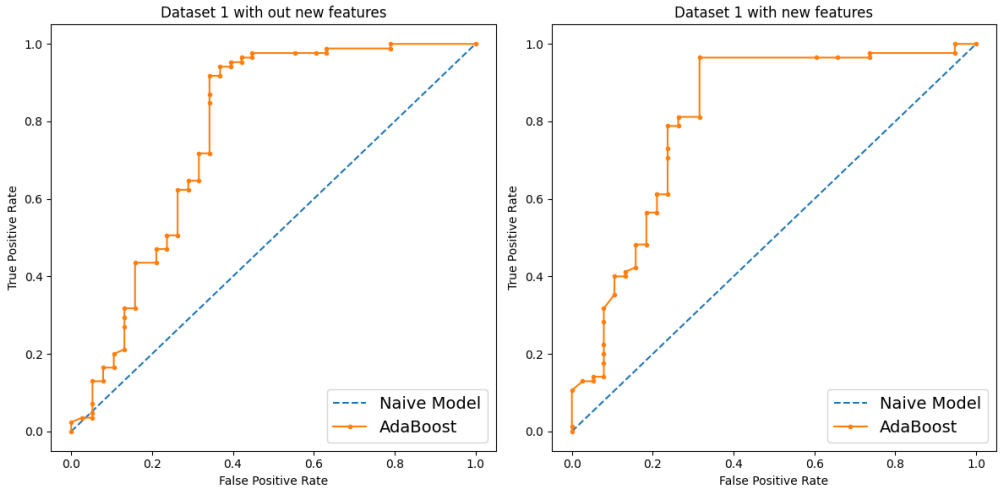

Figure 7. ROC Curves of AdaBoost in the Baseline Model and the Proposed Model

https://doi.org/10.36596/jcse.v5i2.812

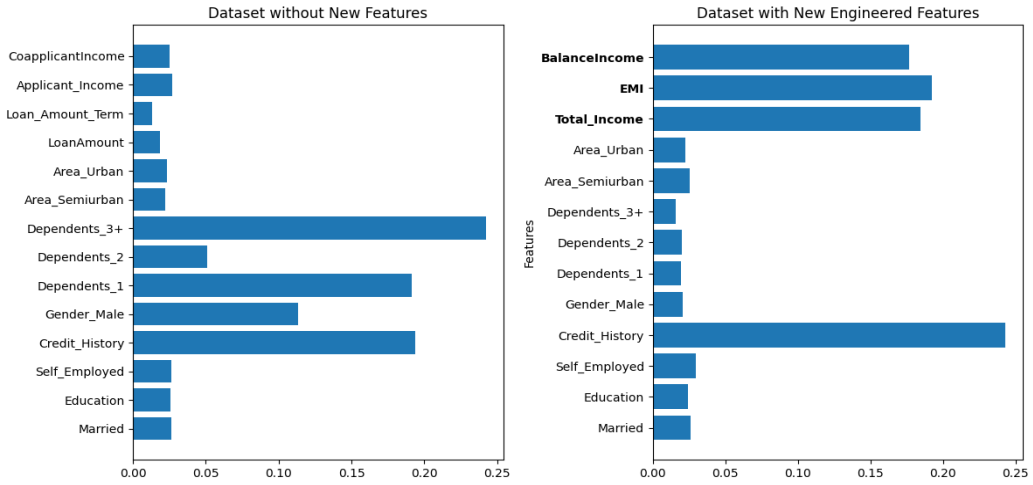

Figure 8. Feature Importance based on Random Forest in the Baseline Model and the Proposed Model

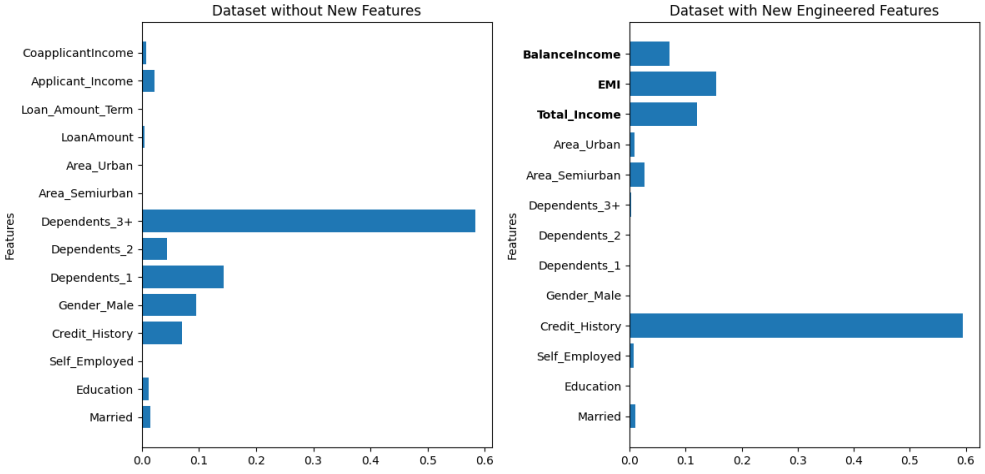

Figure 9. Feature Importance based on Gradient Boosting in the Baseline Model and the Proposed Model

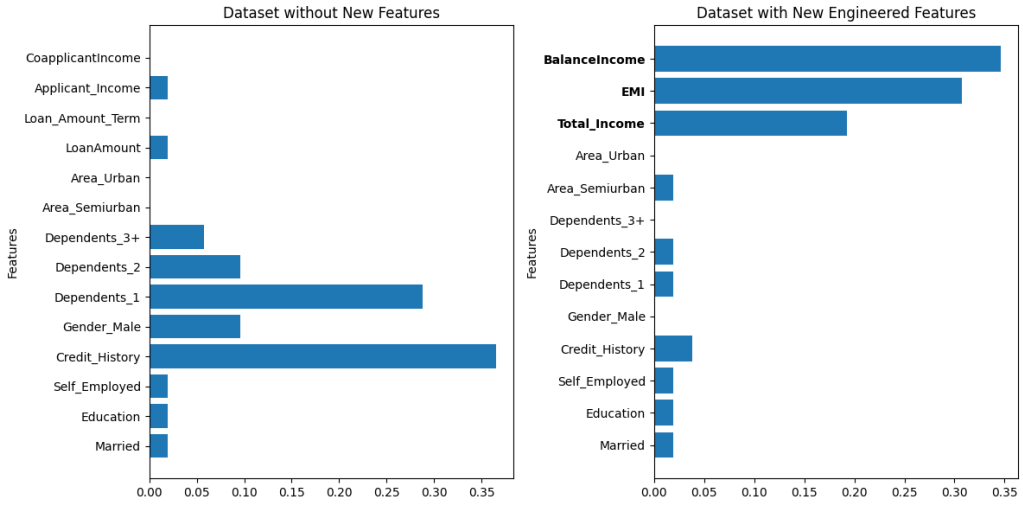

Figure 10. Feature Importance based on AdaBoost in the Baseline Model and the Proposed Model

https://doi.org/10.36596/jcse.v5i2.812

6. Conclusion

In this study, we effectively employed various classification algorithms to predict bank loan approvals, aiming to determine whether a loan can be assigned to a customer based on their financial and behavioral features. Our comprehensive data preprocessing approach, involving cleansing, transformation, and feature engineering, played a pivotal role in enhancing accuracy. In the feature engineering stage, we introduced beneficial financial factors like EMI, Total Income, and Net Income, derived from existing features. Employing seven classifiers, including classic and ensemble methods, on the preprocessed data provided reliable and accurate classification. Besides, a thorough analysis was conducted using diverse performance metrics such as Accuracy, F1 score, Recall, Precision, ROC Area, and Feature Importance. Among these classifiers, AdaBoost emerged as the most accurate with an 88% accuracy rate, while the K-nearest neighbor exhibited the lowest accuracy at 68%.

To emphasize the importance of newly devised features, the ROC curve and feature importance were selected to compare the performance of the preprocessed model without feature engineering (baseline model) and the performance of the proposed model. The larger area under the ROC curve and lower false positive rate for the presented model make it generate more reliable classifications compared to the baseline model. Moreover, feature importance comparison indicated that the baseline model relied more on customer behavior features like 'Dependent_3', which contradicts with low correlation result obtained by the heatmap graph. Conversely, the suggested model offers a more robust and accurate classification by highly relying on financial customer features like EMI, total income, and balance income. This underscores the substantial impact of innovative features introduced during the feature engineering stage. Ultimately, our classification model provides a more robust and accurate basis for loan credit approval by identifying problematic clients among a large pool of loan applicants.

References

[1] M. Anand, A. Velu, and P. Whig, "Prediction of Loan Behaviour with Machine Learning Models for Secure Banking," Journal of Computer Science and Engineering (JCSE), vol. 3, no. 1, pp. 1–13, Feb. 2022, doi: 10.36596/jcse.v3i1.237.

[2] M. Udhbav, R. Kumar, N. Kumar, R. Kumar, Dr. M. Vijarania, and S. Gupta, "Prediction of Home Loan Status Eligibility using Machine Learning," SSRN Electronic Journal. Elsevier BV, 2022. doi: 10.2139/ssrn.4121038.

[3] M. Kumar, V. Goel, T. Jain, S. Singhal and L. Goel, "Neural Network Approach To Loan Default Prediction", International Research Journal of Engineering and Technology (IRJET), vol. 05, 2018, [online] Available: https://www.irjet.net/archives/V5/i4/IRJET-V5I4942.pdf.

[4] R. C. Chiang, Y.-F. Chow, and M. Liu, The Journal of Real Estate Finance and Economics, vol. 25, no. 1. Springer Science and Business Media LLC, pp. 5–32, 2002. doi: 10.1023/a:1015347516812.

[5] M.-C. Chen and S.-H. Huang, "Credit scoring and rejected instances reassigning through evolutionary computation techniques," Expert Systems with Applications, vol. 24, no. 4. Elsevier BV, pp. 433–441, May 2003. doi: 10.1016/s0957-4174(02)00191-4.

[6] T. Sueyoshi, "DEA-discriminant analysis in the view of goal programming," European Journal of Operational Research, vol. 115, no. 3. Elsevier BV, pp. 564–582, Jun. 1999. doi: 10.1016/s0377-2217(98)00014-9..

[7] U. Aslam, H. I. Tariq Aziz, A. Sohail, and N. K. Batcha, "An Empirical Study on Loan Default Prediction Models," Journal of Computational and Theoretical Nanoscience, vol. 16, no. 8. American Scientific Publishers, pp. 3483–3488, Aug. 01, 2019. doi: 10.1166/jctn.2019.8312.

[8] M. Shoaeinaeini, O. Ozturk, and D. Gupta, "Twitter-informed Prediction for Urban Traffic Flow Using Machine Learning," 2022 6th International Conference on Universal Village (UV). IEEE, Oct. 22, 2022. doi: 10.1109/uv56588.2022.10185516.

[9] O. Ozturk, B. Hangun, and M. Shoaeinaeini, "Utilizing Machine Learning to Predict Offshore Wind Farm Power Output for European Countries," 2022 11th International Conference on Renewable Energy Research and Application (ICRERA). IEEE, Sep. 18, 2022. doi: 10.1109/icrera55966.2022.9922823.

[10] J. Wiens and E. S. Shenoy, "Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology," Clinical Infectious Diseases, vol. 66, no. 1. Oxford University Press (OUP), pp. 149–153, Aug. 21, 2017. doi: 10.1093/cid/cix731.

https://doi.org/10.36596/jcse.v5i2.812

[11] H. A. P. L. Perera and S. C. Premaratne, "An Artificial Neural Network Approach for the Predictive Accuracy of Payments of Leasing Customers in Sri Lanka," presented at the International Conference on Business, Economics, and Social Science & Humanities (BESSH-2016), Novotel Hotel Sydney Central, Sydney, Australia, Sep. 2016, vol. 2.
[12] Z. Ereiz, “Predicting Default Loans Using Machine Learning (OptiML),” 2019 27th Telecommunications Forum (TELFOR). IEEE, Nov. 2019. doi: 10.1109/telfor48224.2019.8971110.
[13] Y. Li, “Credit Risk Prediction Based on Machine Learning Methods,” 2019 14th International Conference on Computer Science & Education (ICCSE). IEEE, Aug. 2019. doi: 10.1109/iccse.2019.8845444.
[14] A. Bagherpour, "Predicting Mortgage Loan Default with Machine Learning Methods," University of California/Riverside, 2017.
[15] A. Steenackers and M. J. Goovaerts, “A credit scoring model for personal loans,” Insurance: Mathematics and Economics, vol. 8, no. 1. Elsevier BV, pp. 31–34, Mar. 1989. doi: 10.1016/0167-6687(89)90044-9.
[16] M. Madaan, A. Kumar, C. Keshri, R. Jain, and P. Nagrath, “Loan default prediction using decision trees and random forest: A comparative study,” IOP Conference Series: Materials Science and Engineering, vol. 1022, no. 1. IOP Publishing, p. 012042, Jan. 01, 2021. doi: 10.1088/1757-899x/1022/1/012042.
[17] L. Zhu, D. Qiu, D. Ergu, C. Ying, and K. Liu, “A study on predicting loan default based on the random forest algorithm,” Procedia Computer Science, vol. 162. Elsevier BV, pp. 503–513, 2019. doi: 10.1016/j.procs.2019.12.017.
[18] M. S. Irfan Ahmed and P. Ramila Rajaleximi, "An Empirical Study on Credit Scoring and Credit Scorecard for Financial Institutions," International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 8, no. 7, pp. 275-279, Jul. 2019, ISSN: 2278-1323..
[19] A. Jafar Hamid and T. M. Ahmed, “Developing Prediction Model of Loan Risk in Banks Using Data Mining,” Machine Learning and Applications: An International Journal, vol. 3, no. 1. Academy and Industry Research Collaboration Center (AIRCC), pp. 1–9, Mar. 30, 2016. doi: 10.5121/mlaij.2016.3101.
[20] P. Supriya, M. Pavani, N. Saisushma, N. Vimala Kumari, and K. Vikas, "Loan Prediction by using Machine Learning Models," International Journal of Engineering and Techniques, vol. 5, no. 2, pp. 144-148, Mar.-Apr. 2019.
[21] J. Tejaswini, T. M. Kavya, R. D. N. Ramya, P. S. Triveni, and V. R. Maddumala, "Accurate loan approval prediction based on machine learning approach," Journal of Engineering Science, vol. 11, no. 4, pp. 523-532, 2020.
[22] P. Tumuluru, L. R. Burra, M. Loukya, S. Bhavana, H. M. H. CSaiBaba, and N. Sunanda, “Comparative Analysis of Customer Loan Approval Prediction using Machine Learning Algorithms,” 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS). IEEE, Feb. 23, 2022. doi: 10.1109/icais53314.2022.9742800.
[23] Shoaei, Maryam. “Dream Housing Finance Dataset.” Kaggle, May 18, 2024. https://www.kaggle.com/datasets/maryamshoaei/dream-housing-finance-dataset/data.
[24] L. HOTA, “A Comparative Performance Assessment for Prediction of Loan Approval in Financial Sector.” Research Square Platform LLC, Apr. 04, 2023. doi: 10.21203/rs.3.rs-2763466/v1.
[25] Viswanatha V, Ramachandra A.C, Vishwas K N, and Adithya G, “Prediction of Loan Approval in Banks using Machine Learning Approach”, int. j. eng. mgmt. res., vol. 13, no. 4, pp. 7–19, Aug. 2023.
[26] “Sklearn.Ensemble.Adaboostclassifier.” scikit. Accessed October 20, 2023. https://scikitlearn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html#sklearn.ensemble.AdaBoostClassifier.

https://doi.org/10.36596/jcse.v5i2.812