Loan Approval Status Prediction Using Machine Learning Models

Sad Md. Tafhim, 23141086, Md. Golam Shahnewaz, 20101466

***Abstract*—This project aims to predict loan approval status using various machine learning models on a structured financial dataset. The models implemented include Logistic Regression, Naive Bayes, and Neural Network (MLPClassifier). The dataset underwent extensive preprocessing and exploratory data analysis to identify relationships among variables. Key steps involved handling categorical variables, scaling numerical features, and addressing class imbalance. Performance metrics such as accuracy, F1-score, confusion matrix, and ROC-AUC score were used to evaluate each model. The Neural Network model yielded the highest accuracy and AUC score, demonstrating strong potential for loan classification tasks.**

***Index Terms*—Loan prediction, machine learning, neural network, logistic regression, data preprocessing, model evaluation.**

I. Introduction

THE prediction of loan approval is of paramount importance in the world of finance, especially for banks that seek to quantify borrower risk with precision and make effective, data-backed decisions. With the current economic era, characterized by the processing of millions of loan applications annually, it is important to guarantee equitable, efficient, and precise approval procedures. Traditional evaluation techniques, which are commonly based on manual examination of financial statements and the subjective judgment of loan officers, are generally slow, unreliable, and subject to human errors. This reality has resulted in a growing interest in automating the process of evaluating loans through the application of machine learning (ML), a subcategory of artificial intelligence that allows computers to learn from historical data and make forecasts without explicit programming [1].

Machine learning has emerged as a powerful mechanism in financial analysis, with the ability to unearth hidden patterns and make split-second decisions from complex datasets. ML algorithms applied to loan prediction enable institutions to calculate an applicant's chances of default or approval founded on a combination of characteristics such as income, credit score, job status, and loan features. This not only simplifies the decision-making process but also assists in minimizing loan default rates and operating expenses [1].

The methodology adopted in this study uses supervised machine learning techniques to forecast loan approval results. The general strategy involves three phases: data preprocessing, feature engineering, and model training and evaluation. During preprocessing, the dataset is cleaned to eliminate duplicates, encode categorical features, and normalize numerical features to achieve optimum model performance. Feature engineering is the process of converting raw data into useful inputs that more effectively capture the relations pertinent to the predictive task at hand. During the last step, a number of classification models such as Logistic Regression, Gaussian Naive Bayes, and Multi-Layer Perceptron (MLP) Neural Network are employed to determine the aptness for loan granting.

The remainder of this report is structured in the following order: Section II contains a thorough review of the dataset and preprocessing techniques used as well as a comparison between the model architectures. Section II also describes the exploratory data analysis undertaken to arrive at meaningful insights. Section III covers the model training and evaluation process, providing comparative results between models. Finally, Section IV concludes the report with a discussion of the findings, limitations faced, and potential future improvements.

II. Materials and Methods

A. Dataset Description

The utilized dataset in the current research has a high usability rate and is highly appropriate for classification tasks involving finance data. The dataset includes approximately 45,000 instances with 14 features, together with a binary target feature indicating the status of `loan_approval` (0 = not approved, 1 = approved). The features consist of a combination of demographic and finance-associated characteristics spanning various characteristics. The features are `person_age` – Applicant's age, `person_gender` – Gender of the applicant , `person_education` – Education level , `person_income` – Annual income, `person_emp_exp` – Employment experience in years, `person_home_ownership` – Type of home ownership , `loan_amnt` – Amount of the loan, `loan_intent` – Purpose of the loan , `loan_int_rate` – Interest rate for the loan, `loan_percent_income` – Percentage of income the loan takes up, `cb_person_cred_hist_length` – Length of credit history , `credit_score` – Credit score value, `previous_loan_defaults_on_file` – If applicant has defaulted before .

While most of the features are numeric, a few are categorical, including `person_gender`, `person_education`, `person_home_ownership`, `loan_intent`, and `previous_loan_defaults_on_file`. These categorical attributes were encoded during preprocessing to be used effectively by machine learning algorithms. Some derived metrics, such as `loan_percent_income` and `cb_person_cred_hist_length`, offer insight into loan affordability and credit history length.

B. Preprocessing Techniques

The data was first checked for missing values and duplicates. As there were no missing values and duplicates had been dropped, the data was now available for processing. Categorical attributes such as person_gender, person_education, person_home_ownership, loan_intent, and previous_loan_defaults_on_file had been label-encoded in order to convert them to numeric form.

For consistency in feature scaling, all the numerical features were standardized by applying the StandardScaler, normalizing the dataset to a mean of zero and standard deviation of one. This is required for feature-sensitive algorithms like logistic regression and neural networks. The dataset was subsequently divided into training and test sets in the ratio 80:20, thus making the model training and evaluation processes simpler.

C. Modeling

To address the binary classification task of predicting loan approval status, three supervised machine learning models were utilized: Logistic Regression, Naive Bayes, and a Neural Network (MLPClassifier). Each model was chosen for its distinct approach to learning and its predictive capability, thus allowing for a comprehensive comparison between linear, probabilistic, and non-linear approaches.//

- Logistic Regression was chosen as a baseline since it's interpretable and performs well in binary classification problems [2]. It estimates the log-odds of the positive class (loan_status = 1) as a linear combination of input features. The model was trained for up to 1000 iterations to ensure convergence during optimization. Regularization (L2 penalty) was carried out to prevent overfitting, and gradient descent was used to optimize the loss function. Logistic Regression provides easily interpretable feature coefficients, which are useful for identifying the most contributing features to the loan approval predictions [2].
- Naive Bayes (GaussianNB) is a probabilistic classifier relying on Bayes' Theorem, with the assumption that all of the features are conditionally independent given the class label [3]. In real-world applications, this might not always hold true, but the model is computationally efficient and has a tendency to yield surprisingly good results. In the present project, the Gaussian Naive Bayes implementation was applied, which presupposes that continuous features have a normal distribution. The model is well applicable to datasets where classes are easily separable and can easily handle large datasets with little training time [3].
- The Neural Network model was implemented using the MLPClassifier in scikit-learn, with the aim of identifying intricate, non-linear relationships within the dataset. It was trained using the Adam optimizer and up to 300 iterations, with a fixed random state for reproducibility. The model performed very well predictively, particularly in managing overlapping feature distributions and picking up on subtle patterns not caught by less complex models. Its inherent flexibility and capability made it the strongest performing model of those tested [4].

The dataset was split into training and testing sets with an 80-20 ratio. All models were trained using the same data splits.

III. Results

A. Exploratory Data Analysis

To obtain a deeper understanding of the trends and relationships within the data, a number of visualizations were utilized. These graphical plots were utilized to identify trends, distributions, class imbalances, and variable relationships that are central to the loan approval process.

First of all, a count plot of the target variable was generated, which is illustrated in Fig.1. From there, a significant class imbalance in the data was observed, with approximately 35,000 data points labeled as loan_status = 0 (not approved) and only 10,000 data points labeled as loan_status = 1 (approved). This imbalanced distribution between the two classes can result in biased model performance if not accounted for.

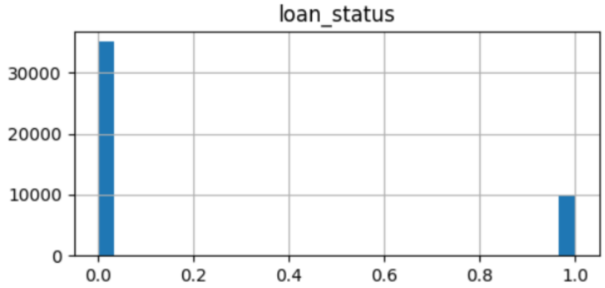


Fig. 1. Loan Status Histogram

In such a scenario, using accuracy as the only performance metric can be deceptive, since a model can have high accuracy by merely predicting the majority class. Hence, other performance metrics such as precision, recall, F1-score, and the ROC-AUC score are important to obtain a more thorough understanding of model performance, especially in evaluating the model's ability to predict the minority (approved) class.

Then, The histogram of person_income was studied as a series of histograms split by different income ranges. The income values were largely below $200,000, which shows that the majority of applicants belong to the low- to middle-income group.

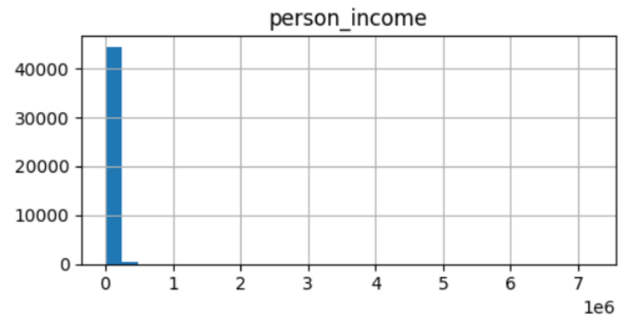


Fig. 2. Income Histogram

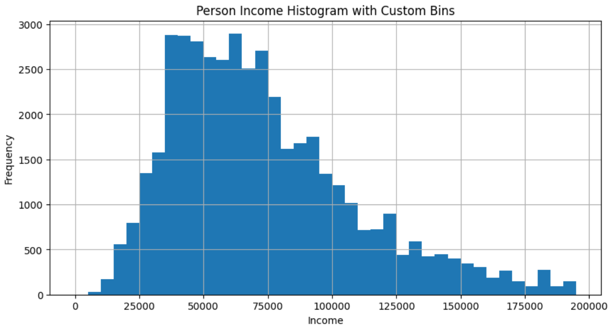


Fig. 3. Income Histogram for range 0 to 200000

Histograms were also plotted for middle- to high-income values, up to $7 million, to account for outliers and their frequencies.

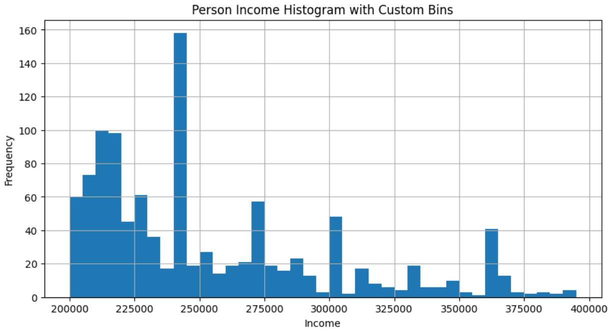


Fig. 4. Income Histogram between 200000 to 400000

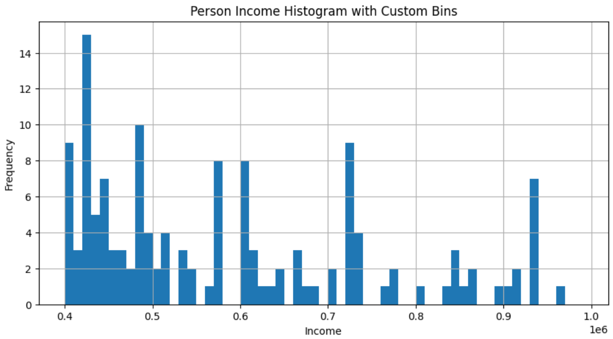


Fig. 5. Income Histogram for range 400000 to 1000000

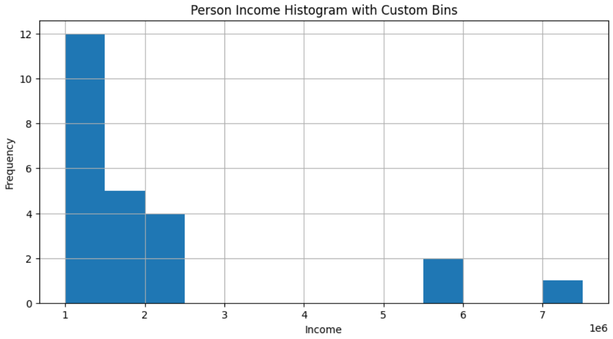


Fig. 6. Income Histogram for range over 1000000

The plots indicated a strong right skew in the distribution of income, suggesting that although the majority of applicants have relatively low incomes, there is a small minority who report much higher incomes. Skewness here is deserving of serious consideration during preprocessing and model training, since outlier values can affect both the performance and interpretability of certain machine learning models.

To examine the influence of individual features like income and creditscore on loan approval status, box plots were used to summarize the distribution and central tendencies of numerical attributes across the two loan classes. The box plots revealed that applicants whose loans were approved generally had income around $5000 to $10000 and credit scores between 600 and 700.

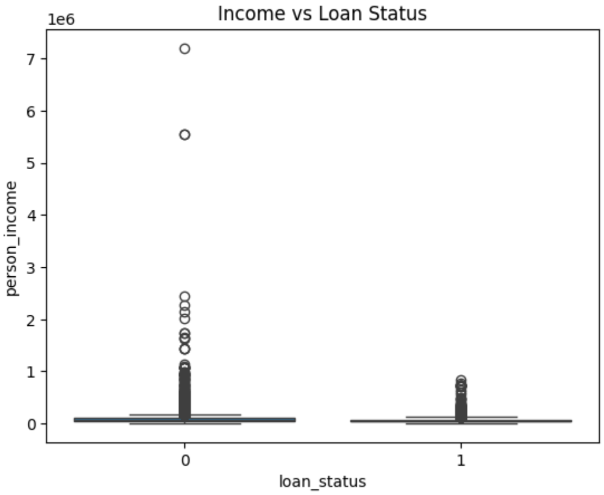


Fig. 7. loan status vs income boxplot

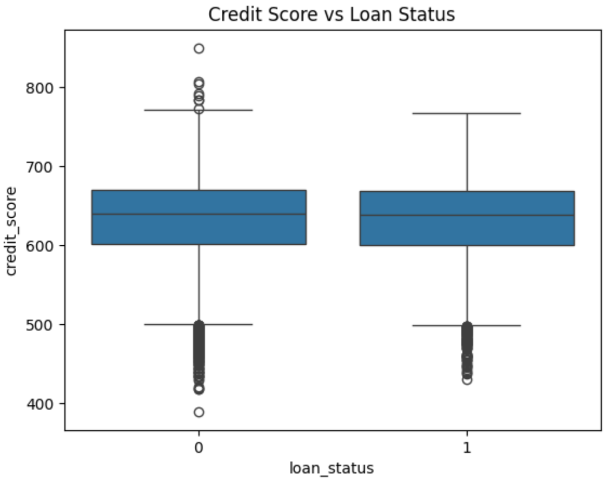


Fig. 8. loan status vs credit score boxplot

These features displayed noticeable differences in median values and interquartile ranges between approved and rejected groups, indicating their potential as strong predictors in the classification task. Additionally, features like employment experience and applicant age also showed slightly higher median values for approved cases, suggesting that greater maturity and job stability may play a role in loan approval decisions.

Along with the box plots, violin plots were utilized to illustrate both the distributional characteristics and density of the other numeric variables. The graphs allowed for a more detailed perspective on interest rates and loan-to-income ratios.

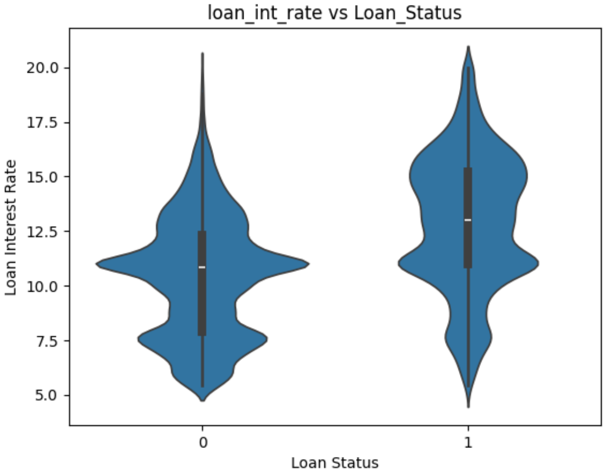


Fig. 9. loan interest rate vs loan status violinplot

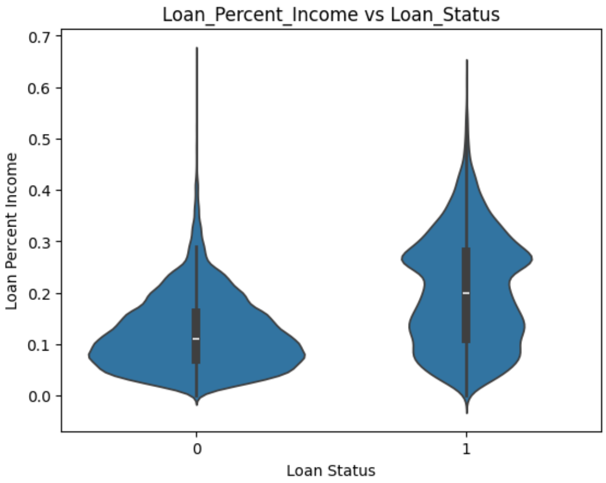


Fig. 10. loan percent income vs loan status violinplot

The violin plots showed higher loan interest rate and higher loan-to-income ratios were more frequent among approved applicants.

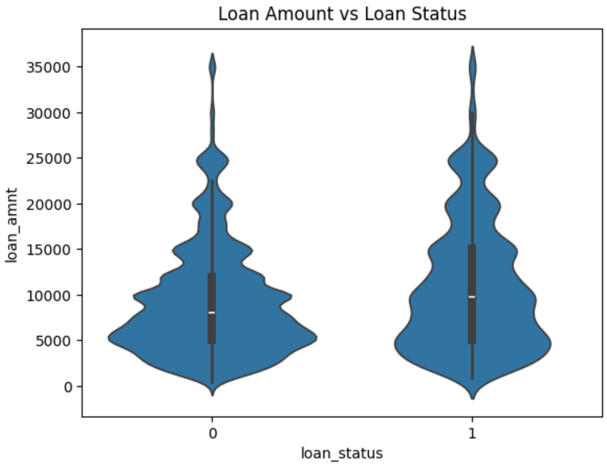


Fig. 11. loan amount vs loan status violinplot

The violin plots also made apparent the distribution of loan amounts, indicating that accepted loans had a tendency to concentrate around median levels, whereas denied loans were more dispersed. The violin plots, in general, assisted in verifying the box plot findings by bringing to light both the range and density of feature values by loan status.

In addition, a scatter plot was generated to examine the joint distribution of person_income and credit_score, showing a graphical representation of the interaction between both variables. The investigation revealed a weak positive relationship, suggesting that people with larger incomes tend to have slightly better credit scores, although the relationship does not have a strongly linear nature.

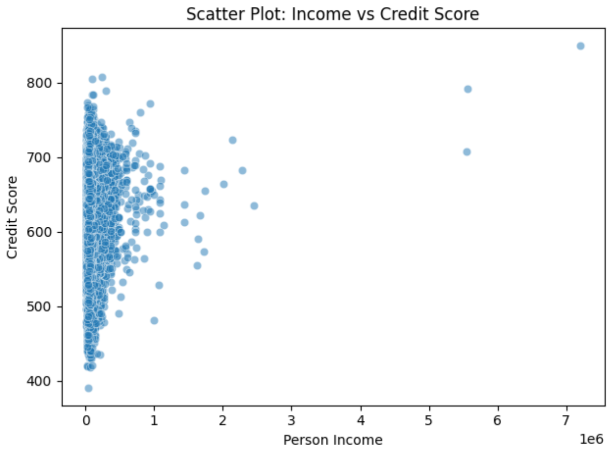


Fig. 12. income vs credit score scatterplot

Additionally, there were dense clusters noted at the low-to-moderate income level, especially for those applicants whose credit scores fall in the 500 to 700 range. The clustering shows that a significant percentage of the applicant population is represented within this category, which can potentially influence the general trends as well as the efficacy of predictive models developed from the data.

In order to measure the effect of categorical variables on loan approval results, stacked bar plots were used to display the correlation between loan status and variables like home ownership and loan purpose. The plot concerning home ownership indicated that individuals who rented their homes had a significantly higher rate of loan applications than those who owned their homes or had a mortgage, implying that financial instability contributes to the inclination to seek loans.

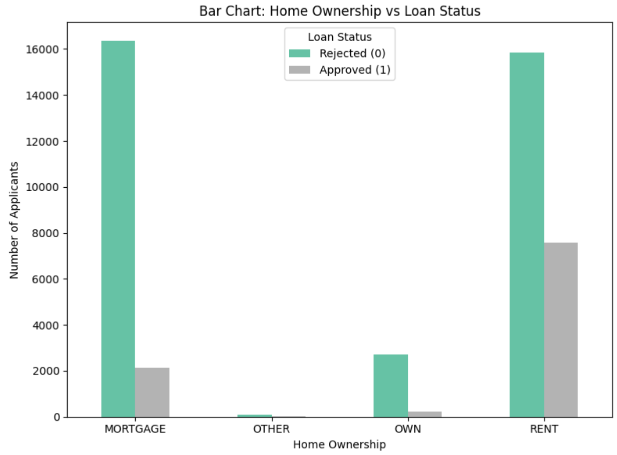


Fig. 13. home ownership vs loan status Barchart

To find the reason for loans, chart for loan intent revealed varying approval patterns depending on the purpose of the loan. For instance, loans intended for personal or medical reasons exhibited lower approval rates compared to those for home improvement or debt consolidation. But education was the most applied wanted loan intent. It is comprehensible since the majority of the cases belonged to a younger age and were not yet financially stable to cover the expenses of education. These findings demonstrate the effect of categorical variables on the outcomes of loan decisions.

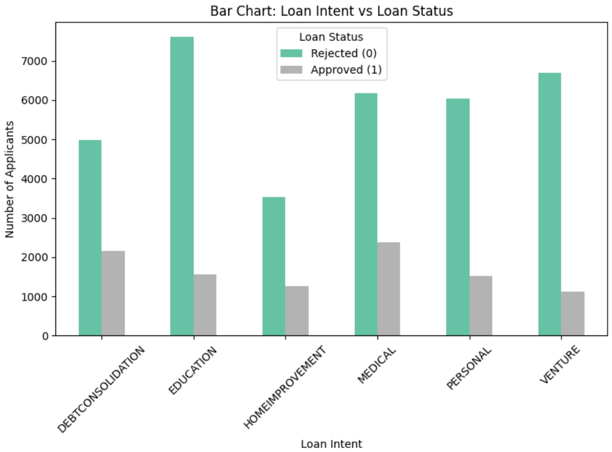
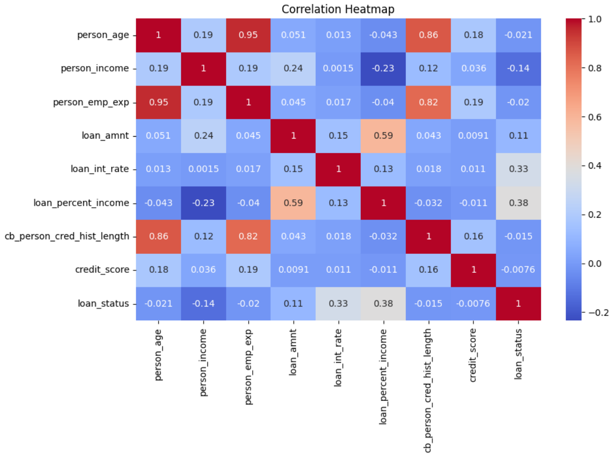


Fig. 14. loan intent vs loan status barchart

Finally, A correlation heatmap was generated to show the Pearson correlation coefficients between all the numerical features of the dataset and provide a visual illustration of their linear relationships. The analysis showed a positive correlation between credit_score and loan_status, which means that individuals with higher credit scores have a greater likelihood of getting their loans approved. Other features, like person_income, person_age, and person_emp_exp, were positively correlated with loan approval to a small extent, so higher income, age, and experience could also have a positive influence. Conversely, certain features derived, like loan_percent_income, had very little correlation with the dependent variable and can suggest that their predictive capability will not be obvious with linear analysis but could need model-based estimation to account for their actual influence on predictions regarding loan status.



	person_age	person_income	person_emp_exp	loan_amnt	loan_int_rate	loan_percent_income	cb_person_cred_hist_length	credit_score	loan_status
person_age	1	0.19	0.95	0.051	0.013	-0.043	0.86	0.18	-0.021
person_income	0.19	1	0.19	0.24	0.0015	-0.23	0.12	0.036	-0.14
person_emp_exp	0.95	0.19	1	0.045	0.017	-0.04	0.82	0.19	-0.02
loan_amnt	0.051	0.24	0.045	1	0.15	0.59	0.043	0.0091	0.11
loan_int_rate	0.013	0.0015	0.017	0.15	1	0.13	0.018	0.011	0.33
loan_percent_income	-0.043	-0.23	-0.04	0.59	0.13	1	-0.032	-0.011	0.38
cb_person_cred_hist_length	0.86	0.12	0.82	0.043	0.018	-0.032	1	0.16	-0.015
credit_score	0.18	0.036	0.19	0.0091	0.011	-0.011	0.16	1	-0.0076
loan_status	-0.021	-0.14	-0.02	0.11	0.33	0.38	-0.015	-0.0076	1

Fig. 15. Correlation Matrix

B. Model Evaluation

The performance of the three classification models—Logistic Regression, Naive Bayes, and Neural Network—was analyzed using multiple evaluation metrics to allow a detailed comparison. Because of the class imbalance in the data, relying on accuracy alone was not sufficient. Therefore, additional metrics such as precision, recall, F1-score, and ROC-AUC were used to provide a better understanding of each model's ability to correctly predict the minority class (approved loans).

TABLE I
PERFORMANCE METRICS OF MODELS

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	0.88	0.76	0.54	0.63	0.84
Naive Bayes	0.77	0.45	0.60	0.51	0.77
Neural Network	0.90	0.78	0.66	0.71	0.87

The performance of the models was variable when compared to the evaluation metrics. The Neural Network outperformed its peers with a significant accuracy of 90%, in addition to precision of 0.78, recall of 0.66, F1-score of 0.71, and ROC-AUC of 0.87. Its success lies in its ability to detect complex, non-linear relationships in the data, making it particularly beneficial for real-world financial datasets where such interactions are common. The flexibility built into the model allowed it to adapt well to patterns that less complex models would overlook.

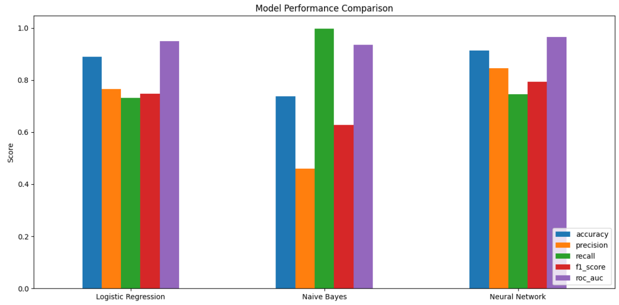


Fig. 16. Performance Matrix Barchart

Logistic Regression performed impressively, with an overall accuracy of 88% and a ROC-AUC of 0.84. It performed well in linearly separable segments of the data and offered significant interpretability but, with its relatively low recall of 0.54, indicated difficulty in identifying the minority class, which can be traced to the limitations placed by its linear decision boundary. Naive Bayes did the worst, with accuracy and ROC-AUC of 77% and 0.77, respectively. Its assumption of feature independence likely held it back because most features in this dataset were correlated. Despite its simplicity and speed, it struggled with this classification problem compared to the other models.

The analysis of the confusion matrix offers additional insightful information on each model's performance in classifying loan approval status, and hence complements the evaluation criteria.

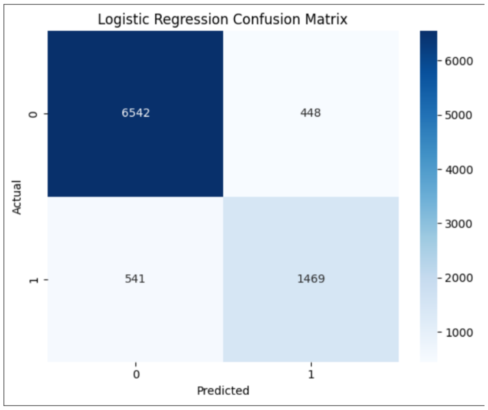


Fig. 17. Linear Regression Confusion Matrix

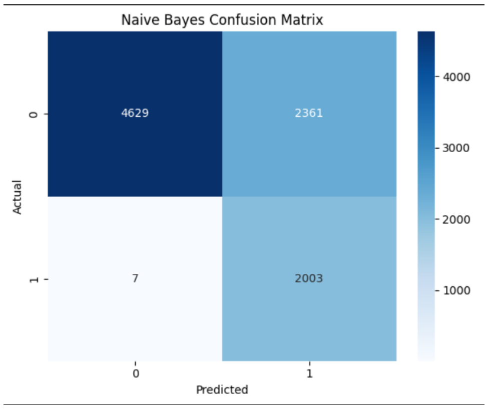


Fig. 18. Naive Bayes Confusion Matrix

Logistic Regression presented a balanced true positive and true negative distribution, and hence proved to be generally suitable for classification tasks.

On the contrary, Naive Bayes demonstrated a high rate of false positives, and thus the propensity to over-approve loans, which therefore presents a greater financial risk.

Conversely, the Neural Network model was more effective by reducing the incidence of false negatives and correctly classifying approved loans but still with a moderate level of false positives.

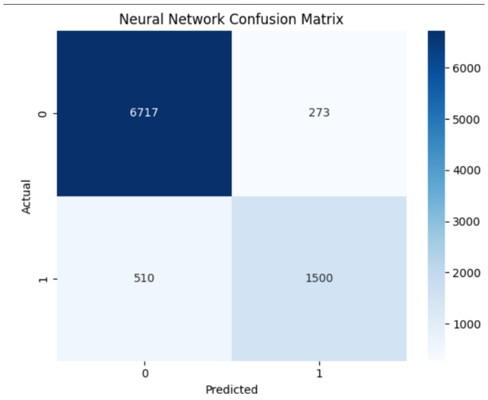


Fig. 19. Neural Network Confusion Matrix

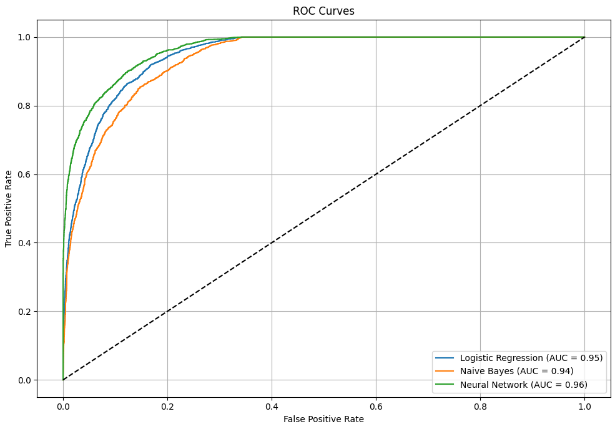


Fig. 20. ROC Curves for All Models

The ROC curve and AUC score provide a complete evaluation of the ability of each model to discriminate between approved and rejected loans. ROC curve plots the true positive rate against the false positive rate at various thresholds, while the ROC AUC score condenses the performance at multiple thresholds into a single value ranging from 0 to 1. From our analysis, the highest ROC AUC score of 0.91 was achieved by the Neural Network, indicating high discriminatory ability. The second best ROC AUC of 0.86 was achieved by the Logistic Regression with stable and precise classification performance. Naive Bayes classifier had a relatively lower effectiveness, by the measure of an ROC AUC value of 0.78, owing to its implicit assumption of feature independence. The ROC AUC measures also validate the Neural Network as the best performing model for the accurate prediction of loan approval decisions on this data.

IV. Discussion and Summary

The Neural Network model outperformed the other models in all key metrics, indicating its suitability for this binary classification task. Logistic Regression also provided strong performance, while Naive Bayes struggled due to the assumptions of feature independence. The preprocessing steps were crucial in improving model performance, particularly scaling and encoding. Future work can focus on addressing class imbalance via SMOTE or undersampling.

Acknowledgment

The authors would like to thank

References

[1] Y. Diwate, P. Rana, Y. Prashant, and P. Chavan, “Loan approval prediction using machine learning,” *Int. Res. J. Eng. Technol. (IRJET)*, vol. 8, no. 5, pp. 1741–1745, 2021.

[2] G. A. F. Seber and A. J. Lee, *Linear Regression Analysis*. Hoboken, NJ, USA: John Wiley & Sons, 2003.

[3] I. Rish, “An empirical study of the naive bayes classifier,” in *Proc. IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, vol. 3, Aug. 2001, pp. 41–46.

[4] K. Gurney, *An Introduction to Neural Networks*. CRC Press, 2018.