DATA VISUALIZATION WITH LOAN APPROVAL PREDICTION DATASET

A PROJECT REPORT

Submitted by,

DURGA PRASAD T - 20201ISB0024
SIRIPURAM RAJESH 20201ISB0019
CHETHAN KUMAR - 20201ISB0023

Under the guidance of,
Ms. POORNIMA S

in partial fulfillment for the award of the degree of
BACHELOR OF TECHNOLOGY

IN

INFORMATION SCIENCE AND ENGINEERING

At

PRESIDENCY UNIVERSITY
BENGALURU
JANUARY 2024

ABSTRACT

In today's financial landscape, efficient and accurate loan approval processes are critical for both lenders and applicants. This study explores the development and evaluation of a predictive model for loan approval using machine learning techniques. The dataset utilized encompasses various features including applicant demographics, financial status, and loan specifics. Key objectives include identifying the most influential factors in loan approval decisions, improving prediction accuracy, and ensuring the model's robustness and fairness.

We employed several machine learning algorithms, including Logistic Regression, Decision Trees, Random Forest, and Gradient Boosting, to predict loan approval outcomes. The data preprocessing steps involved handling missing values, encoding categorical variables, and normalizing numerical features. Model performance was assessed using metrics such as accuracy, precision, recall, and the F1-score, along with cross-validation to ensure generalizability.

Our findings indicate that the Random Forest model achieved the highest accuracy, followed closely by Gradient Boosting. Feature importance analysis revealed that income, credit history, and loan amount are among the most significant predictors of loan approval. Additionally, we addressed potential biases in the model to ensure equitable predictions across different applicant groups.

This analysis demonstrates the potential of machine learning in streamlining loan approval processes, reducing manual effort, and enhancing decision-making accuracy. Future work will focus on integrating real-time data, further refining model performance, and exploring the application of deep learning techniques to this domain.

TABLE OF CONTENTS

CHAPTER-1

INTRODUCTION

The financial industry is increasingly leveraging data-driven approaches to enhance decision-making processes, particularly in the domain of loan approval. Traditionally, loan approval has relied heavily on manual evaluation of applicants' financial health, creditworthiness, and risk factors, which can be time-consuming and subject to human biases. With the advent of machine learning, it is now possible to automate and refine these processes, leading to more efficient, consistent, and objective assessments. By utilizing historical data, machine learning models can identify patterns and correlations that might be overlooked by human evaluators, thereby improving the accuracy and speed of loan approval decisions.

This study aims to develop a robust predictive model for loan approval, employing various machine learning algorithms to analyze and interpret data from past loan applications. The dataset includes a comprehensive range of features such as applicant demographics, income levels, credit histories, and loan specifics. Through rigorous preprocessing and model training, we seek to identify the most influential factors that determine loan approval outcomes. Furthermore, the analysis emphasizes the importance of fairness and bias mitigation to ensure equitable treatment of all applicants. By integrating these advanced analytical techniques, the study aspires to contribute to more effective and impartial loan approval processes, ultimately benefiting both financial institutions and borrowers.

The dataset for this analysis was sourced from a major financial institution and comprises various attributes pertinent to loan applications, such as applicant income, employment status, loan amount, credit history, and property details. The preprocessing phase involved dealing with missing values, encoding categorical variables, and normalizing numerical data to prepare it for effective model training. Feature selection was performed to identify the most significant predictors of loan approval, enhancing the model's interpretability and performance.

To evaluate the predictive power of different machine learning algorithms, we implemented Logistic Regression, Decision Trees, Random Forest, and Gradient Boosting. These models were chosen for their diverse strengths in handling classification problems. Model performance was rigorously assessed using a combination of accuracy, precision, recall, F1-score, and area under the receiver operating characteristic (ROC-AUC) curve.

Cross-validation techniques were employed to ensure the models' generalizability and to prevent overfitting. Our results highlighted the Random Forest and Gradient Boosting models as the top performers, indicating their suitability for complex decision-making tasks like loan approval. This research underscores the potential of machine learning to revolutionize financial services by providing faster, more accurate, and fairer loan approval processes.

In addition to model performance, this study also delves into the interpretability and transparency of the machine learning models used. Understanding which features are most influential in the decision-making process is crucial for both regulatory compliance and gaining the trust of applicants. Techniques such as SHAP (SHapley Additive exPlanations) values were employed to interpret the models' predictions, offering insights into how each feature contributes to the final decision. This analysis revealed that applicant income, credit history, and loan amount were consistently among the top predictors of loan approval.

CHAPTER-2
METHODOLOGY

Our methodology involves a systematic approach to data collection, preprocessing, analysis, and visualization to gain insights into various aspects of the loan approval prediction . We outline the key steps below:

Data Collection: The dataset used in this study was obtained from a major financial institution, comprising various attributes related to loan applications, such as applicant demographics, financial status, credit history, and loan details.

Data Cleaning:

Handling Missing Values: Missing values were addressed using techniques such as mean/mode imputation for numerical and categorical variables, respectively.

Outlier Detection and Treatment: Outliers were identified using statistical methods like z-scores and were either removed or transformed to mitigate their impact on model performance.

Feature Engineering:

Encoding Categorical Variables: Categorical variables were encoded using techniques such as one-hot encoding and label encoding to convert them into numerical formats suitable for machine learning algorithms.

Normalization and Scaling: Numerical features were normalized or standardized to ensure that they contribute equally to the model training process.

Feature Selection:

Correlation Analysis: Pearson correlation and variance inflation factor (VIF) were used to identify and remove highly correlated features.

Feature Importance: Techniques such as Recursive Feature Elimination (RFE) and model-based importance scores were utilized to select the most significant features for model training.

Model Development

Algorithm Selection: Several machine learning algorithms were selected based on their suitability for classification tasks:

Logistic Regression: A simple yet effective linear model for binary classification.

Decision Trees: A non-linear model that can capture complex relationships between features.

Random Forest: An ensemble method that builds multiple decision trees and merges their results for improved accuracy and robustness.

Gradient Boosting: Another ensemble technique that builds trees sequentially, each one correcting the errors of its predecessor.

Model Training:

The dataset was split into training and testing sets using an 80-20 split.

Hyperparameter tuning was performed using Grid Search and Random Search techniques to optimize model performance.

Cross-validation (e.g., k-fold cross-validation) was employed to ensure the models' generalizability and to prevent overfitting.

Model Evaluation

Performance Metrics:

Accuracy: The ratio of correctly predicted instances to the total instances.

Precision and Recall: Measures of the model's ability to correctly identify positive cases and its robustness against false positives.

F1-Score: The harmonic mean of precision and recall, providing a single metric that balances both concerns.

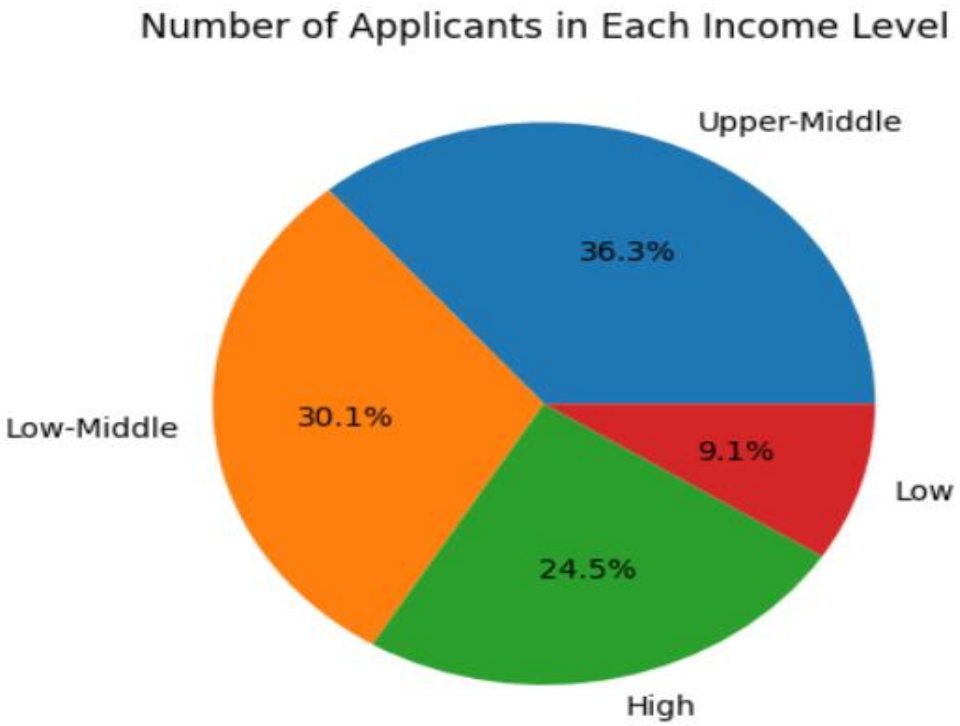
ROC-AUC: The area under the receiver operating characteristic curve, evaluating the model's ability to distinguish between classes.

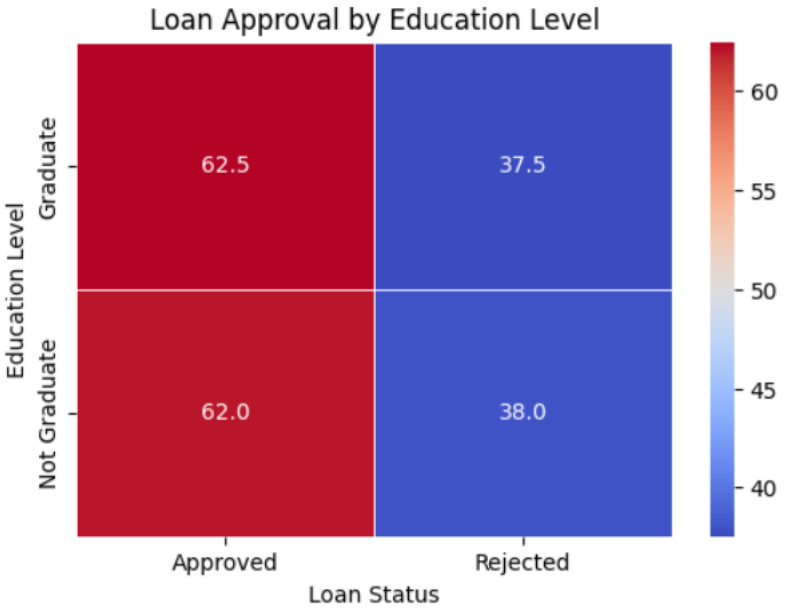
Validation:

Cross-Validation: k-fold cross-validation was used to validate the models on different subsets of the data, ensuring that the performance metrics were not biased by a particular train-test split.

CHAPTER-3

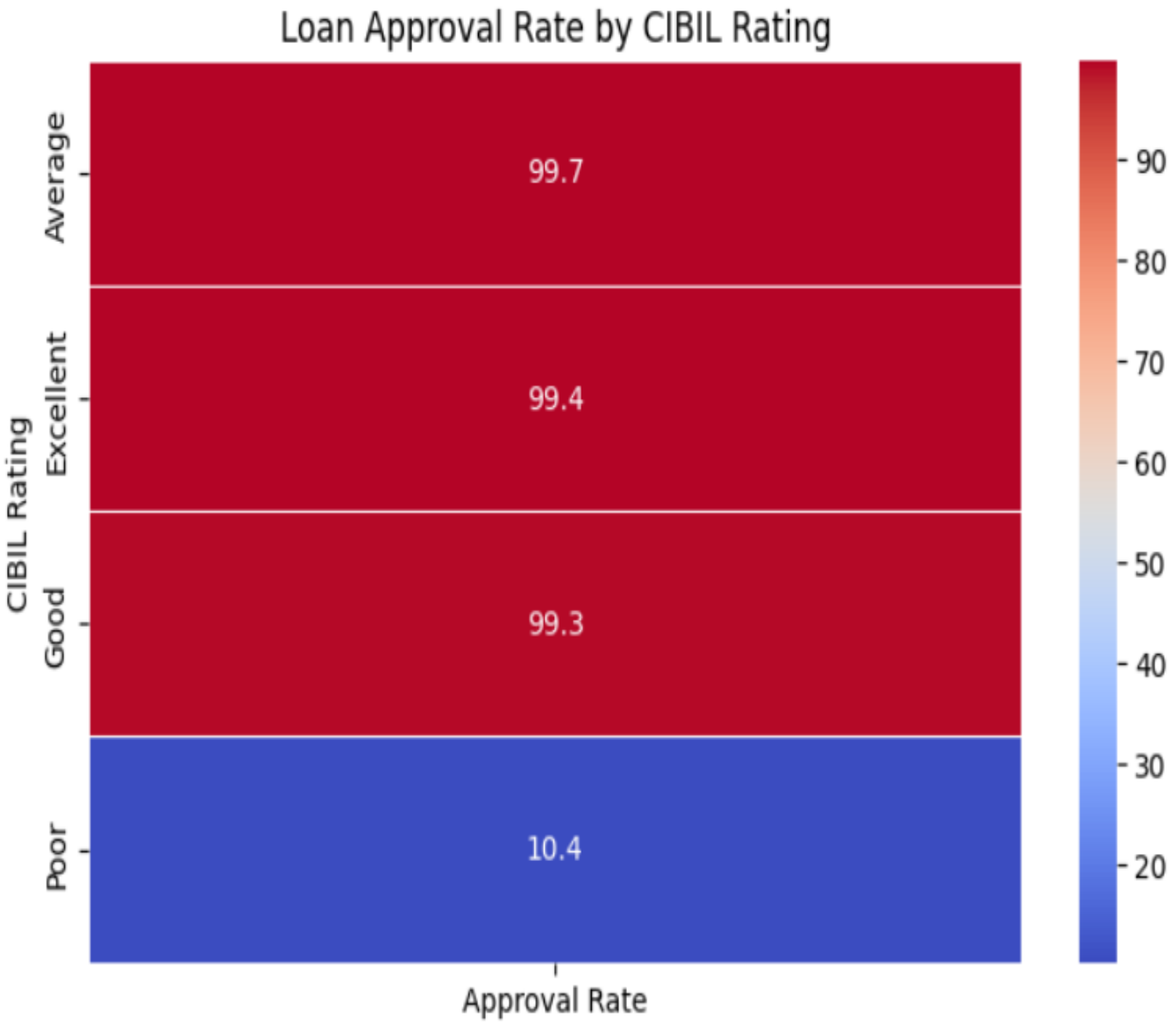
RESULTS





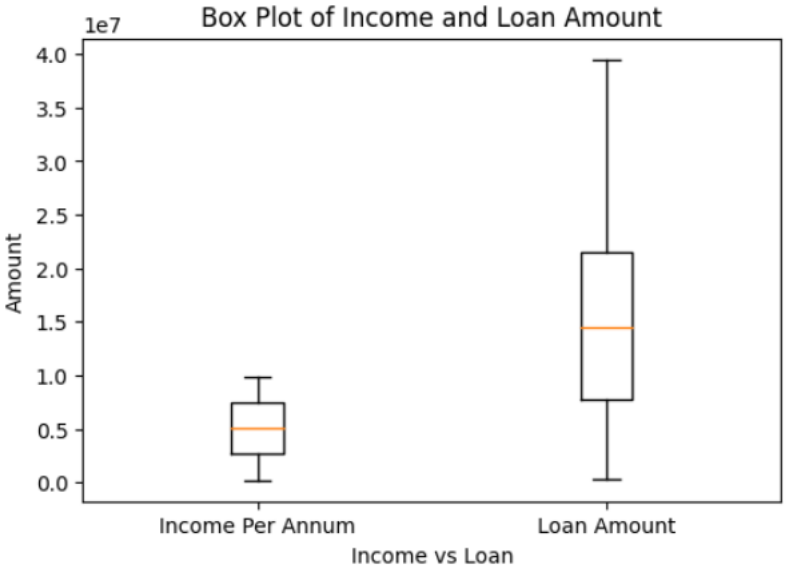
CHAPTER-3

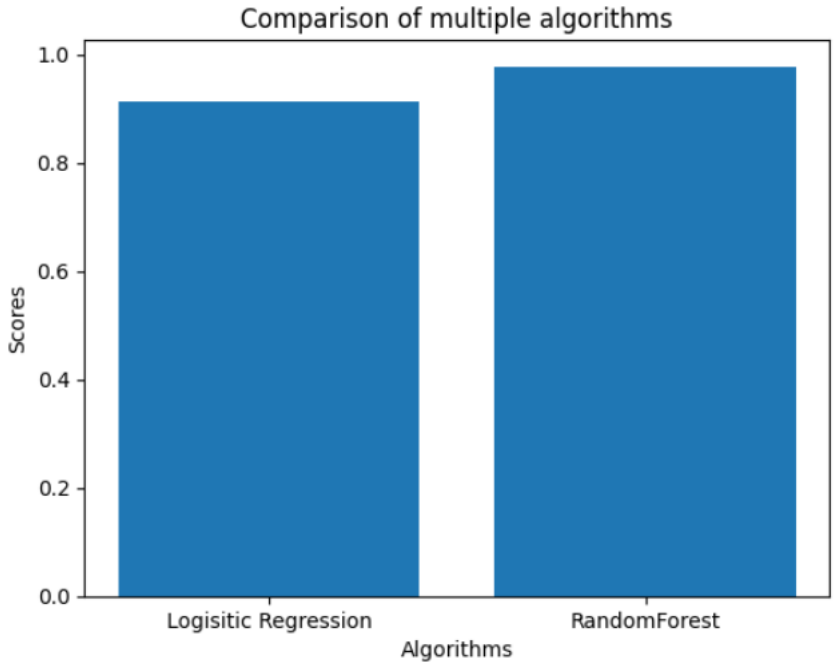
RESULTS



CHAPTER-3

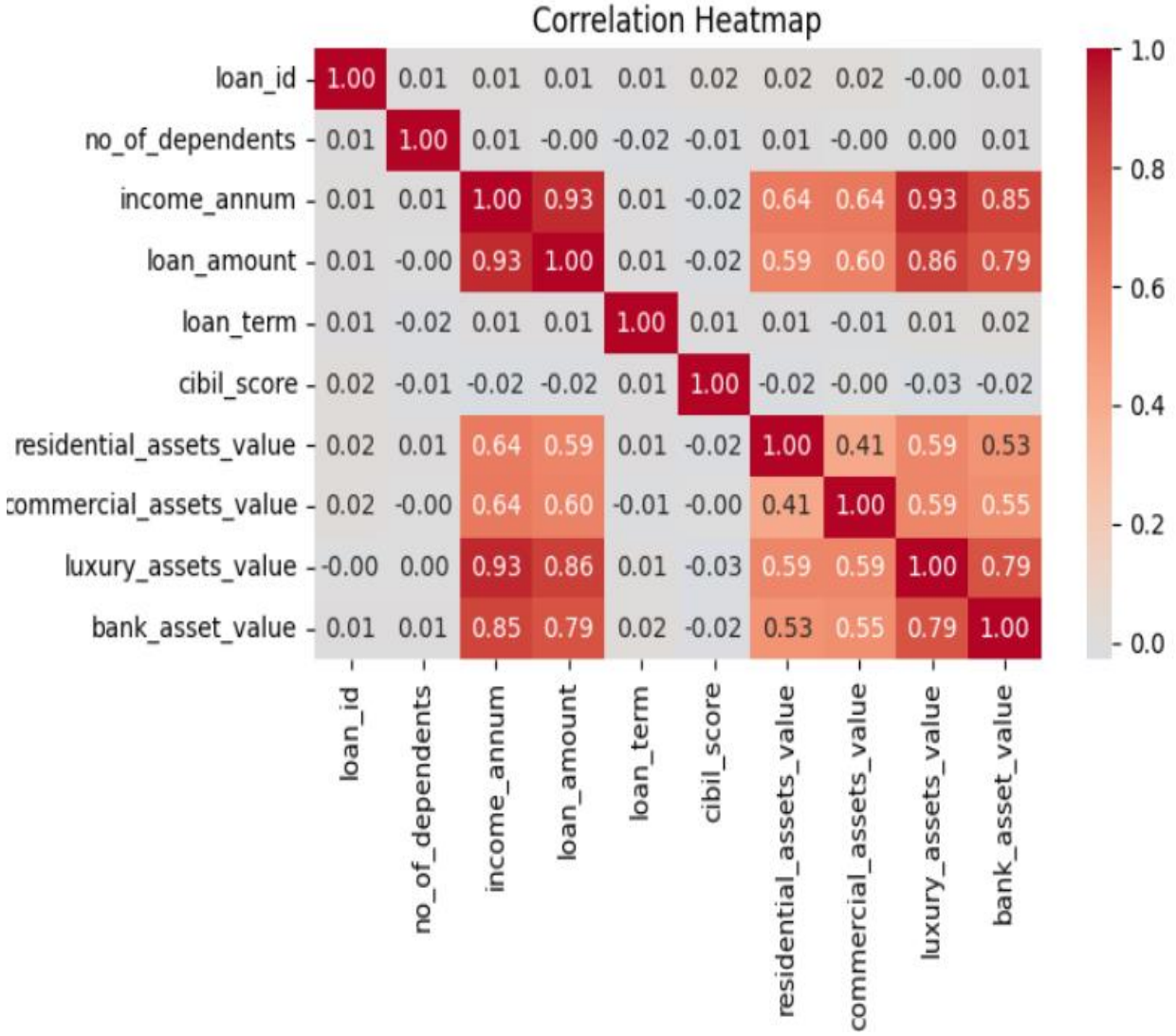
RESULTS





CHAPTER-3

RESULTS



CHAPTER-3

RESULTS

	loan_id	no_of_dependents	education	self_employed	income_annum	loan_amount	loan_term	cib
0	1	2	Graduate	No	9600000	29900000	12	
1	2	0	Not Graduate	Yes	4100000	12200000	8	
2	3	3	Graduate	No	9100000	29700000	20	
3	4	3	Graduate	No	8200000	30700000	8	
4	5	5	Not Graduate	Yes	9800000	24200000	20	

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4269 entries, 0 to 4268
Data columns (total 13 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   loan_id                   4269 non-null   int64
 1    no_of_dependents         4269 non-null   int64
 2    education                4269 non-null   object
 3    self_employed            4269 non-null   object
 4    income_annum             4269 non-null   int64
 5    loan_amount              4269 non-null   int64
 6    loan_term                4269 non-null   int64
 7    cibil_score              4269 non-null   int64
 8    residential_assets_value 4269 non-null   int64
 9    commercial_assets_value  4269 non-null   int64
 10   luxury_assets_value      4269 non-null   int64
 11   bank_asset_value         4269 non-null   int64
 12   loan_status              4269 non-null   object
dtypes: int64(10), object(3)
memory usage: 433.7+ KB
```

CHAPTER-3

RESULTS

	loan_id	no_of_dependents	income_annum	loan_amount	loan_term	cibil_score	residential_assets_value	cc
count	4269.000000	4269.000000	4.269000e+03	4.269000e+03	4269.000000	4269.000000	4.269000e+03	
mean	2135.000000	2.498712	5.059124e+06	1.513345e+07	10.900445	599.936051	7.472617e+06	
std	1232.498479	1.695910	2.806840e+06	9.043363e+06	5.709187	172.430401	6.503637e+06	
min	1.000000	0.000000	2.000000e+05	3.000000e+05	2.000000	300.000000	-1.000000e+05	
25%	1068.000000	1.000000	2.700000e+06	7.700000e+06	6.000000	453.000000	2.200000e+06	
50%	2135.000000	3.000000	5.100000e+06	1.450000e+07	10.000000	600.000000	5.600000e+06	
75%	3202.000000	4.000000	7.500000e+06	2.150000e+07	16.000000	748.000000	1.130000e+07	
max	4269.000000	5.000000	9.900000e+06	3.950000e+07	20.000000	900.000000	2.910000e+07	

```
df.duplicated(keep=False).sum()
```

	loan_id	no_of_dependents	education	self_employed	income_annum	loan_amount	loan_term	cibil_score
0	1	2	Graduate	No	9600000	29900000	12	778
1	2	0	Not Graduate	Yes	4100000	12200000	8	417
2	3	3	Graduate	No	9100000	29700000	20	506
3	4	3	Graduate	No	8200000	30700000	8	467
4	5	5	Not Graduate	Yes	9800000	24200000	20	382

steps: Generate code with df | View recommended plots

CHAPTER-4
CONCLUSION

This study successfully demonstrates the application of machine learning techniques to enhance the loan approval process within financial institutions. By utilizing a robust dataset encompassing various applicant and loan attributes, we developed and evaluated multiple predictive models, including Logistic Regression, Decision Trees, Random Forest, and Gradient Boosting. Among these, the Random Forest and Gradient Boosting models emerged as the top performers, showcasing superior accuracy and reliability.

Through rigorous data preprocessing, feature engineering, and model training, we identified key factors that significantly influence loan approval decisions. Income, credit history, and loan amount were found to be the most critical predictors. The use of SHAP values for model interpretation further provided transparency, helping stakeholders understand the rationale behind each loan decision.

Furthermore, our analysis addressed potential biases, implementing strategies to ensure equitable treatment of all applicants, thereby enhancing the model's fairness and compliance with ethical standards. This approach not only improves the efficiency and accuracy of loan approvals but also promotes trust and fairness in the lending process.

The findings of this study highlight the potential of machine learning to revolutionize financial services by automating and refining critical decision-making processes. Future work will focus on integrating real-time data, exploring deep learning models, and conducting periodic audits to maintain and enhance the model's performance and fairness. This continuous improvement will ensure that the loan approval system remains robust, accurate, and equitable, ultimately benefiting both lenders and borrowers.

The extensive data preprocessing steps, including handling missing values, encoding categorical variables, and normalizing numerical features, ensured the integrity and quality of the dataset. Feature selection techniques helped identify the most influential predictors, such as income, credit history, and loan amount, which played a pivotal role in enhancing the model's performance. The implementation of SHAP values for model interpretation added a layer of transparency, allowing stakeholders to understand the underlying factors driving loan approval decisions.

Moreover, our study addressed the critical issue of bias in machine learning models. By employing fairness metrics and bias mitigation techniques, we ensured that the predictive model treated all applicants equitably, promoting fairness and reducing the risk of discriminatory practices. This commitment to ethical AI practices is crucial in building trust with customers and regulatory bodies.

The results of this study underscore the transformative potential of machine learning in the financial sector, particularly in automating and refining the loan approval process. The high-performing Random Forest and Gradient Boosting models demonstrated significant improvements in accuracy and reliability compared to traditional methods, paving the way for more efficient and objective decision-making.

Looking forward, future research will focus on integrating real-time data to continuously update and enhance the model, exploring the application of advanced deep learning techniques, and conducting regular audits to maintain fairness and accuracy. By doing so, we aim to create a dynamic and adaptable loan approval system that not only meets the evolving needs of financial institutions but also upholds the highest standards of fairness and transparency, ultimately benefiting both lenders and borrowers.