

Risk Analysis for Home Credit Default: Exploratory Data Analysis and Predictive Modeling

PROJECT REPORT

Submitted by

- **Akash Narayan P(Team Lead)**
- **Harikrishnan.S**
- **Yandrathi Tejaswini**
- **Jayaram Gidituri**
- **Kabilan R**
- **Shikhar Pandey**

Under the guidance of

MENTOR:Narendra Kumar

INFOSYS SPRINGBOARD INTERNSHIP 5.0

ABSTRACT

The "Risk Analysis for Home Credit Default: Exploratory Data Analysis and Predictive Modeling" project addresses the critical challenge of predicting loan defaults in the home credit sector. By analyzing a comprehensive Kaggle dataset containing demographic, financial, and loan-specific features, this project aims to identify key factors influencing credit default and develop machine learning models to assess default risk. The project involves performing exploratory data analysis (EDA) to uncover patterns and relationships within the data, followed by feature engineering to enhance model performance.

Various machine learning algorithms, including logistic regression, XGBoost, random forests and neural networks, are trained and evaluated to predict the likelihood of default for home credit applicants. The performance of these models is assessed using standard evaluation metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. The results provide valuable insights into credit risk management and can be utilized by financial institutions to make informed lending decisions. The final deliverable includes the deployment of the predictive models as real-time risk assessment tools, seamlessly integrating into loan origination processes, and providing actionable recommendations to mitigate credit default risk. This project demonstrates the power of machine learning in optimizing decision-making and reducing financial risk in lending.

Table of Contents

1. Introduction.....	1
2. Project Vision and Key goals.....	2
3. Understanding the Data.....	3
4. Data Setup and Initial Exploration.....	4
5. Exploratory Data Analysis.....	5
6. Feature Engineering.....	12
6.1 Payment History (200 Points).....	12
6.2 Credit Utilization (150 Points).....	13
6.3 Credit History Length (100 Points).....	13
6.4 Employment Stability (100 Points).....	13
6.5 Loan Grade and Interest (100 Points).....	13
7. Model Development and Training.....	15
7.1 Random Forest Model.....	15
7.2 Logistic Regression.....	15
7.3 Neural Networks.....	16
7.4 XGBoost.....	16
8. Development and Functionality of the Application.....	18
9. Literature Survey.....	19
10. Output Screens.....	20
10.1 User Interface.....	20
11. Conclusion.....	22
12. Future Scope.....	23

Table of Figures

Figure 1 : A) Distribution of customers across different age groups	
B) Percentage of defaulters for each category of age group.....	5
Figure 2 :A) Distribution of customers across various income categories	
B) Percentage of defaulters for each category of income.....	5
Figure 3 :A) Distribution of customers across each type of Home Ownership	
B) Percentage of defaulters for each category of Home Ownership.....	6
Figure 4 : A) Distribution of customers across various employment duration	
B) Percentage of defaulters by employment durations.....	7
Figure 5 : A) Distribution of customers for each type of Loan Intent	
B) Percentage of defaulters by Loan Intent category	7
Figure 6 :A)Distribution of Loan Amount	
B) Percentage of defaulters by loan amount range.....	8
Figure 7 :A) Distribution of Loan Term Years	
B) Percentage of defaulters by Loan Term Year.....	9
Figure 8 :A) Percentage of defaulters for Each Historical Default Category	
B) percentage of defaulters for each credit history length category.....	9
Figure 9 : A) Percentage of defaulters by Loan Interest Rate Category	
B) Distribution of Current Loan Status.....	10
Figure 10 : A) Correlation Heatmap of Numerical Features	
B) Correlation Heatmap of Categorical Features.....	11
Figure 11: Table with Comparison and Analysis of different classifier models.....	16
Figure 12:ROC Curve Comparison Across Classifier Models.....	17
Figure 13 :Index page with description of features.....	20

Figure 14 :Index page with input details.....	20
Figure 15 :Prediction Result after submitting input details.....	21

1.Introduction

In the Home Credit Default Risk Analysis project, we use machine learning to help financial institutions make better decisions when giving loans. By studying a detailed dataset with information about applicants' personal details, financial history, and loan records, we aim to create models that can predict whether someone is likely to repay a loan or not. This helps reduce the risk for lenders while ensuring borrowers are treated fairly and given equal opportunities to access credit.

We start by carefully studying a dataset from Kaggle, which is the foundation of our project. This involves checking how the data is organized and finding any issues like missing values or unusual data points that could cause problems for our models. Fixing these issues early helps us prepare the data for the next steps.

A key part of our project is feature engineering, where we look closely at the data to find useful patterns and relationships between different pieces of information. We also create new features and change some data into forms that work better for machine learning. This helps our models make more accurate predictions.

For model building, we use popular machine learning methods like Random Forest, XGBoost, Logistic Regression, and Neural Networks to predict if someone might default on a loan. Each model is trained, tested, and compared to find the best one. We also explain how these models work and which factors are most important in making predictions.

Our project isn't just about technical tools and algorithms it's about solving a real-world problem. Financial institutions often face challenges in balancing risks and opportunities. Our solution aims to make loan decisions more reliable, reducing bad debts while ensuring fair treatment for borrowers.

This project provides hands-on experience with machine learning while addressing a real-world challenge. Its insights aim to improve how loans are assessed, approved, and managed for fairer and more reliable decisions.

2. Project Vision and Key Goals

This project uses machine learning to assist financial institutions in making better loan decisions by predicting the likelihood of credit default. By analyzing applicant information such as demographics, financial history, and loan details, we aim to create predictive models that are both accurate and fair.

- **Dataset Exploration and Preprocessing:**

We begin by exploring a Kaggle dataset, identifying missing values, and addressing inconsistencies. This ensures the data is clean and ready for further processing, such as feature engineering and model development.

- **Exploratory Data Analysis (EDA):**

EDA helps uncover patterns and relationships within the dataset. Key distributions and correlations are visualized to guide feature selection and improve the overall quality of the models.

- **Feature Engineering:**

New features are created, and categorical variables are transformed into numerical formats for compatibility with machine learning algorithms.

- **Model Development:**

Predictive models like Random Forest, XGBoost, Logistic Regression, and Neural Networks are trained to identify borrowers at risk of loan default.

- **Model Evaluation and Comparison:**

Model performance is evaluated using metrics like accuracy, precision, recall, F1-score, and AUC-ROC. Different algorithms are compared to identify the best-performing model for predicting default risks.

- **Applications in Financial Institutions:**

These models can help financial institutions improve loan approval processes, reduce risks, and ensure fair access to credit for borrowers

3. Understanding the Data

To predict the likelihood of default by different borrowers, we can utilize a range of historical and transactional data from the various provided datasets. The dataset used here is sourced from [Kaggle](#) . Each file offers critical information that can enhance our understanding of borrower behavior, financial history, and creditworthiness. Here's how each dataset contributes to the default prediction:

- **customer_id**: Unique identifier for each customer.
- **customer_age**: Age of the customer.
- **customer_income**: Income of the customer.
- **employment_duration**: Length of employment in months.
- **loan_amnt**: Amount of loan requested.
- **loan_int_rate**: Interest rate of the loan.
- **term_years**: Duration of the loan in years.
- **historical_default**: Whether the customer has defaulted previously (Y, N, or NaN).
- **cred_hist_length**: Length of the customer's credit history in years.
- **home_ownership_Other**: Whether the customer owns a home or not.
- **loan_intent**: Purpose of the loan (e.g., Education, Medical, Personal).
- **loan_grade**: Grade assigned to the loan (B, C, D, etc.).
- **Current_loan_status**: Status indicating if the customer defaulted.

The dataset combines demographic, financial, and transactional details to predict loan default risk, assessing stability and repayment burden through features like age, income, and loan terms. Historical data and variables such as credit history, loan intent, and current status reveal borrower behavior for accurate predictions.

4. Data Setup and Initial Exploration

The dataset consists of 28,369 rows and 13 columns, representing a detailed collection of borrower information. One of the key steps in preparing the data for machine learning models is addressing missing values, as they can negatively impact model performance and accuracy.

We employed a structured approach to handle the missing values, based on the data types of the respective columns. The following methods were applied:

- **Numerical Columns:** Missing values in numerical columns were filled using the median of the respective column. The median is chosen because it is less sensitive to outliers than the mean, making it more suitable for datasets where extreme values may skew the average. By filling missing values with the median, we preserve the central tendency of the data without introducing bias due to extreme values.
- **Categorical Columns:** For categorical columns, missing values were filled with the most frequent value or the mode of the column. The mode represents the most common category within the data, making it a reasonable replacement for missing values in categorical features. This approach helps maintain the distribution of categories and ensures that no information is lost in the data.

After applying these methods, all missing values in the dataset were successfully filled, and the dataset was made complete and ready for further analysis. With the missing data addressed, we can now proceed with exploratory data analysis, feature engineering, and model development, ensuring that the machine learning models will be trained on a clean and consistent dataset.

Carefully handling missing values is crucial for preparing the dataset for predictive modeling, ensuring accuracy and consistency. This process prevents biases, leading to more reliable models and accurate predictions of loan default risk.

5. Exploratory Data Analysis

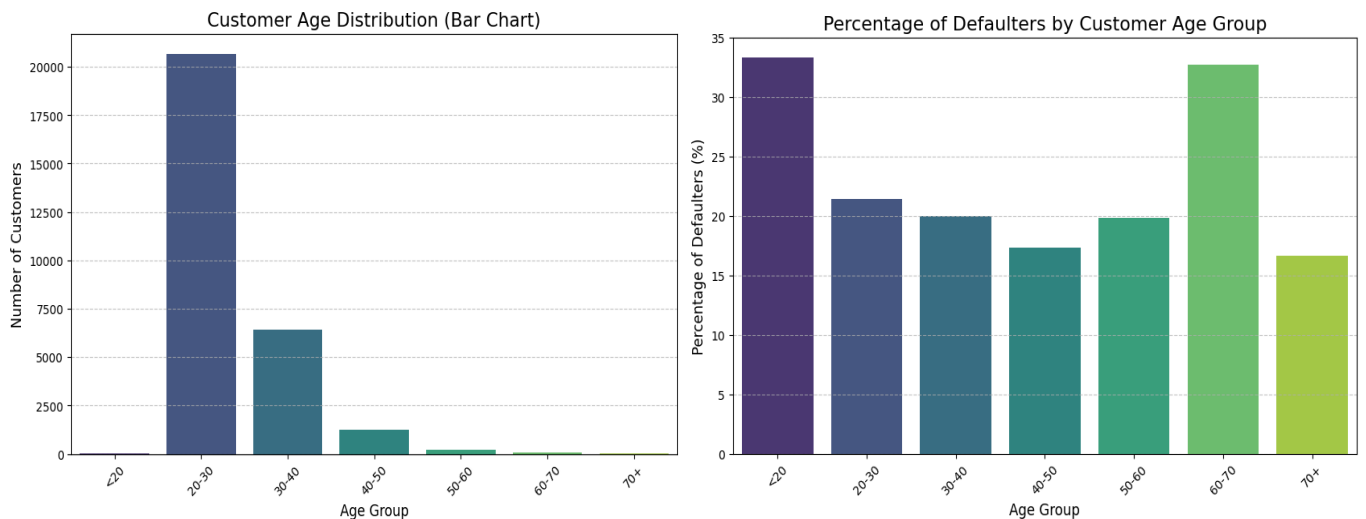


Figure 1: A) Distribution of customers across different age groups B) Percentage of defaulters for each category of age group

The first bar chart depicts the distribution of customers across various age groups, showcasing the number of customers in each range.

The second bar chart illustrates the percentage of loan defaulters across different customer age groups, highlighting age-related risk trends. Customers under the age of 20 tend to default the most, while those over 70 tend to default the least.

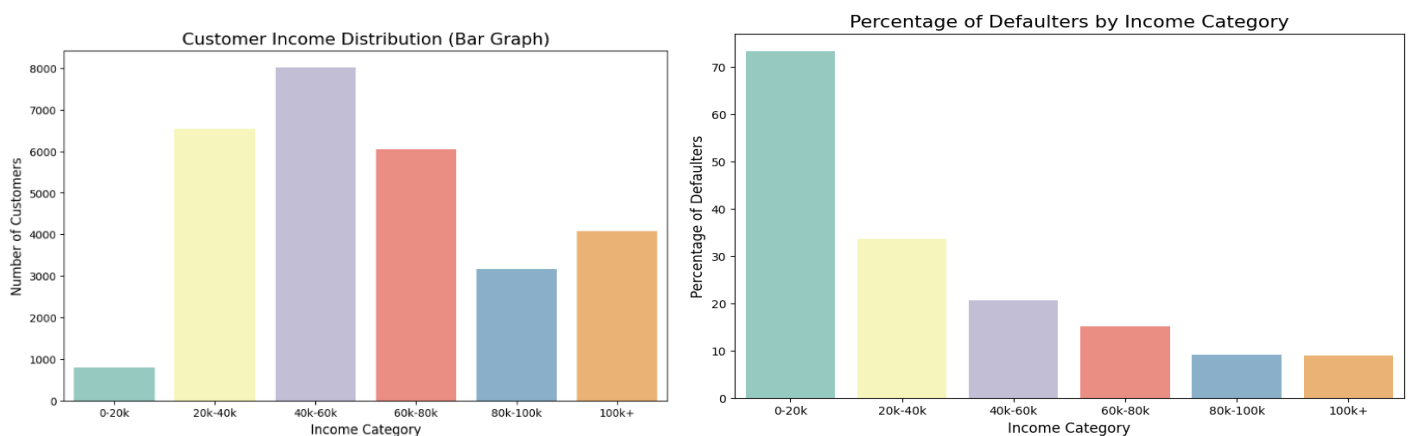


Figure 2: A) Distribution of customers across various income categories B) Percentage of defaulters for each category of income

The first bar chart showing the distribution of customers across income categories helps in understanding how customer income is distributed in the dataset.

The second bar chart displaying the percentage of defaulters across income categories helps in identifying income groups with higher default rates. The customers with an income ranging from 0k to 20k tend to default the most, while those with an income over 100k tend to default the least.

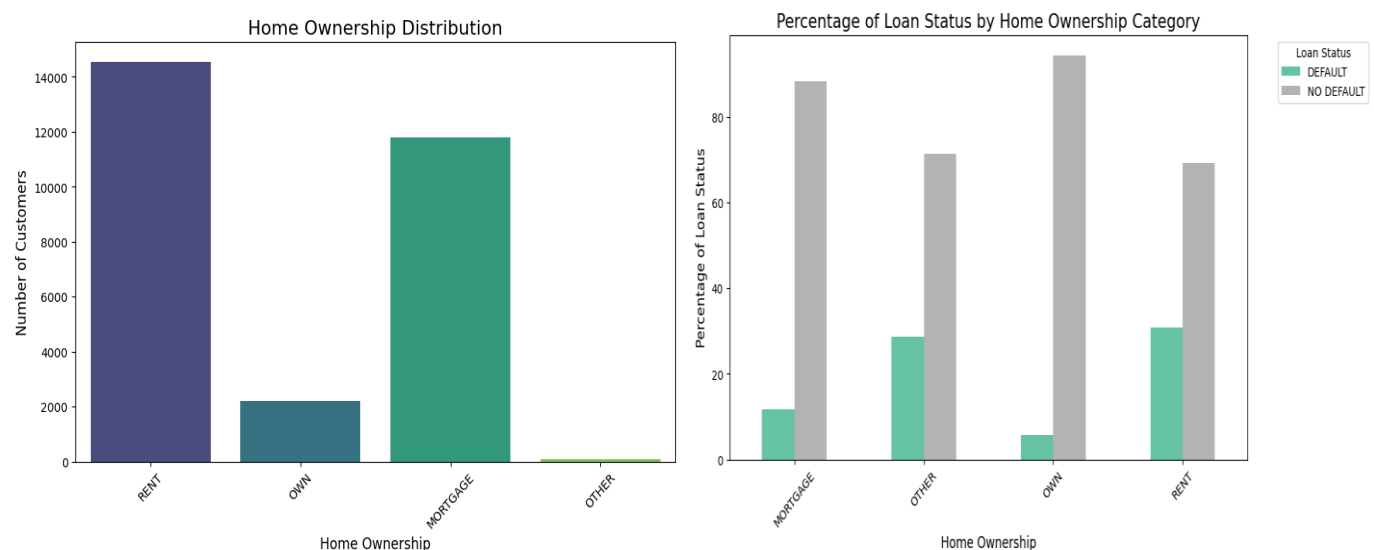


Figure 3: A) Distribution of customers across each type of Home Ownership B) Percentage of defaulters for each category of Home Ownership

The first bargraph helps to understand the distribution of home ownership categories in the dataset, providing insights into the proportion of customers with different home ownership statuses.

The second graph helps to examine the relationship between home ownership status and loan repayment behavior by showing the percentage of defaulters and non-defaulters within each home ownership category. The customers who own a home tend to default the least, while those who rent tend to default the most.

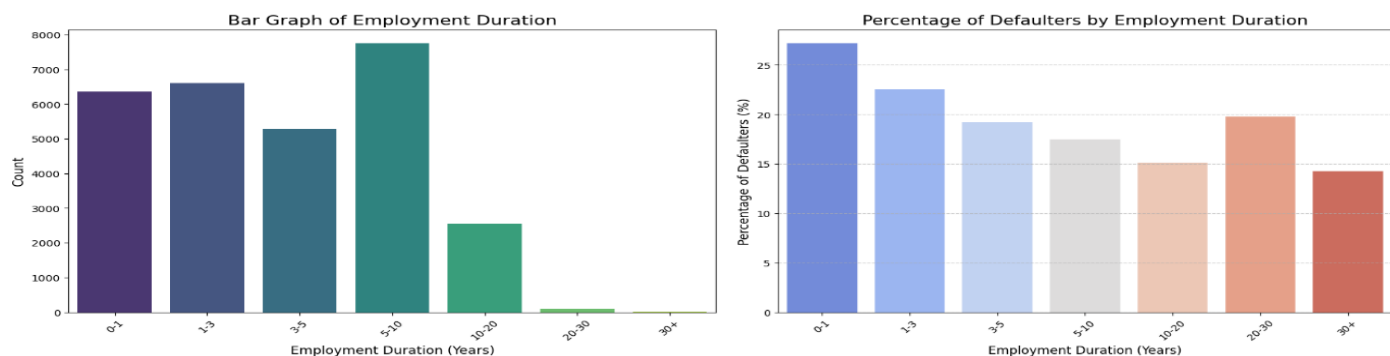


Figure 4 : A) Distribution of customers across various employment duration B) Percentage of Defaulter by employment durations

The first bargraph helps in understanding the distribution of employment durations across customers, which can provide insights into job stability and its potential correlation with loan default risk.

The second bargraph illustrates a negative correlation between employment duration and loan default risk. Borrowers with longer employment durations tend to have a lower likelihood of defaulting on loans, while higher default rates are observed among those with shorter employment histories. This trend suggests that stable employment plays a significant role in reducing loan default risk.

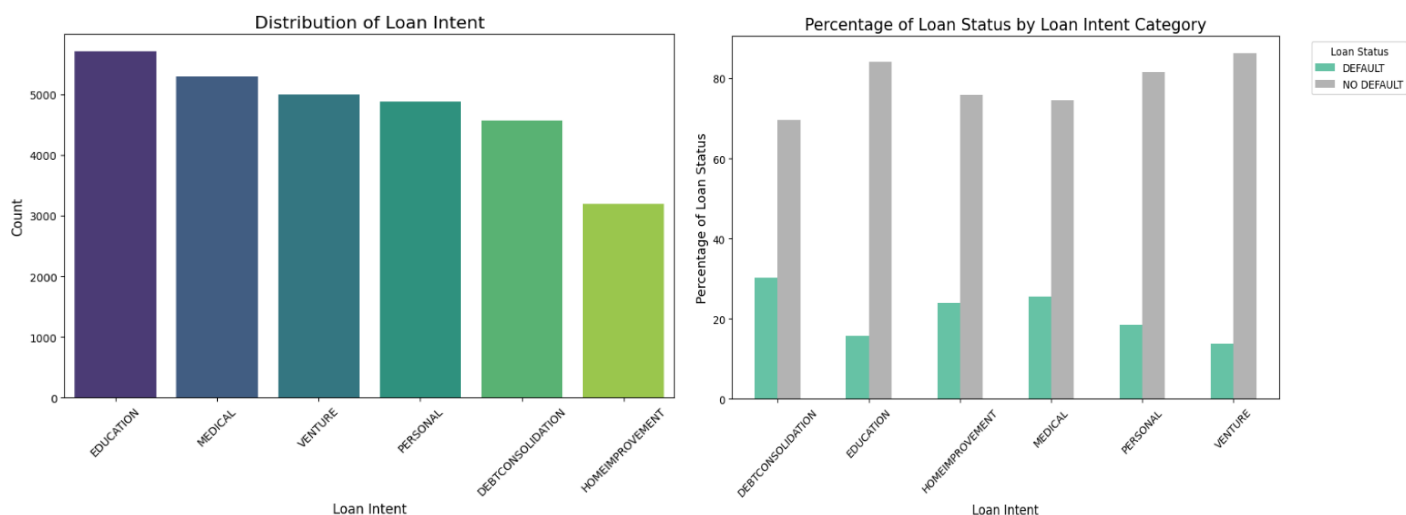


Figure:5 A) Distribution of customers for each type of Loan Intent B) Percentage of defaulters by Loan Intent category

The first bargraph visualizes the distribution of loan intents in the dataset, helping to understand the frequency of different loan purposes and their potential impact on the analysis.

The second bargraph reveals that customers taking loans for ventures and education are the most consistent in repayment, indicating their lower default risk. In contrast, those borrowing for debt consolidation and medical expenses show the highest default rates, highlighting the increased financial strain associated with these purposes

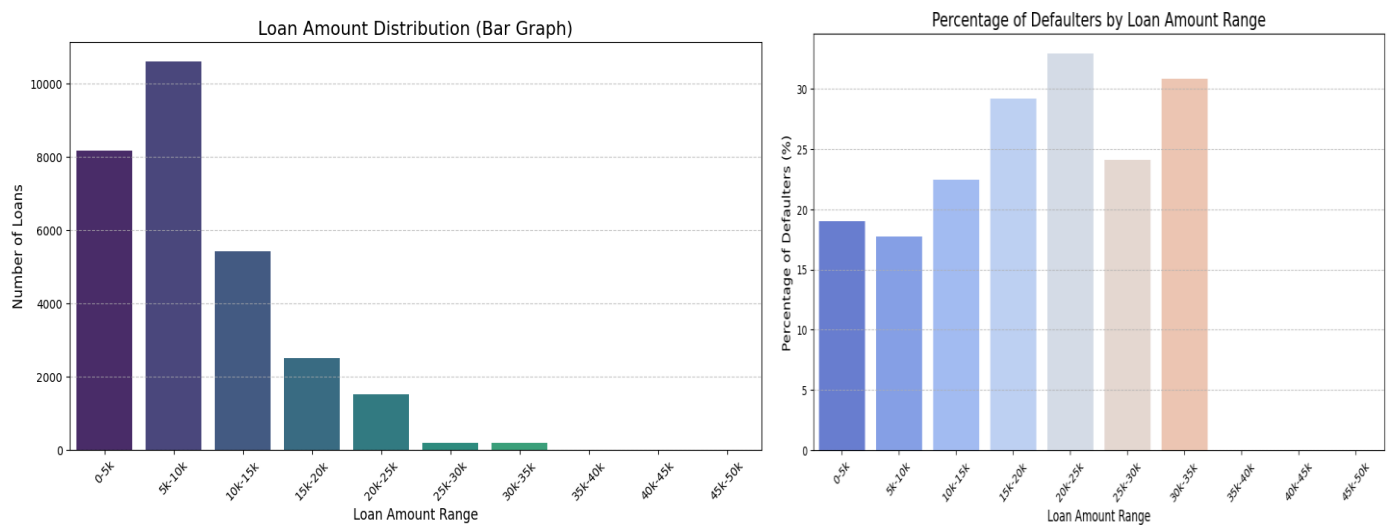


Figure 6 : A)Distribution of Loan Amount B) Percentage of defaulters by loan amount range

The first bargraph helps identify how default rates vary across different loan amount ranges, providing insights into which loan amounts are more likely to result in defaults.

The second bargraph indicates that customers with larger loan amounts (above \$20K) exhibit a lower default percentage, suggesting a higher financial stability or repayment capacity. Conversely, smaller loans (below \$20K) are associated with relatively higher default rates, potentially reflecting greater financial vulnerability among these borrowers.

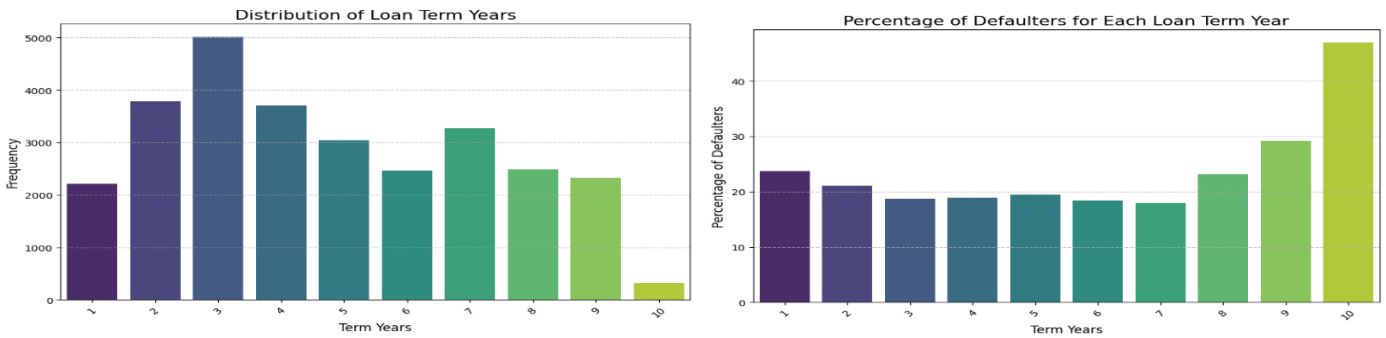


Figure 7: A)Distribution of Loan Term Years B) Percentage of defaulters by Loan Term Year

The first bargraph shows the analysis that helps in EDA by revealing the distribution of loan term durations, aiding in understanding borrower preferences and potential default trends linked to loan term lengths.

The second bargraph reveals that customers with shorter loan terms (1-5 years) tend to have a lower default percentage, indicating better repayment consistency. In contrast, longer loan terms (9-10 years) are associated with higher default rates, possibly due to prolonged financial commitment and increased uncertainty over time.

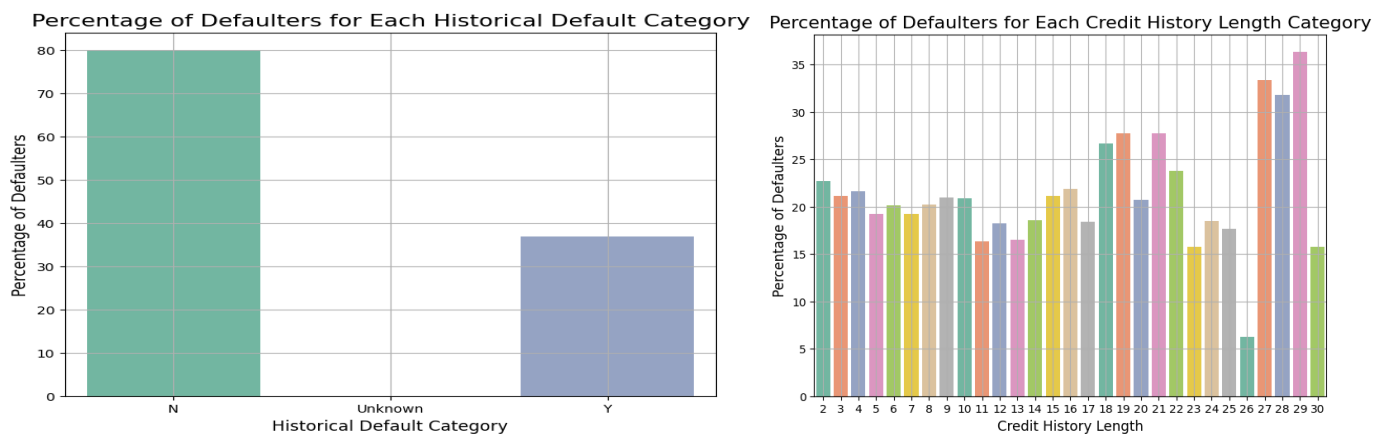


Figure 8: A) Percentage of defaulters for Each Historical Default Category B)percentage of defaulters for each credit history length category

In the first graph, the analysis is useful in EDA as it uncovers the relationship between historical default patterns and the likelihood of loan default, providing insights into how past credit behavior impacts current risk.

In the second graph, the analysis helps in EDA by revealing how the length of credit history influences the likelihood of loan default, aiding in risk assessment based on borrower credit experience.

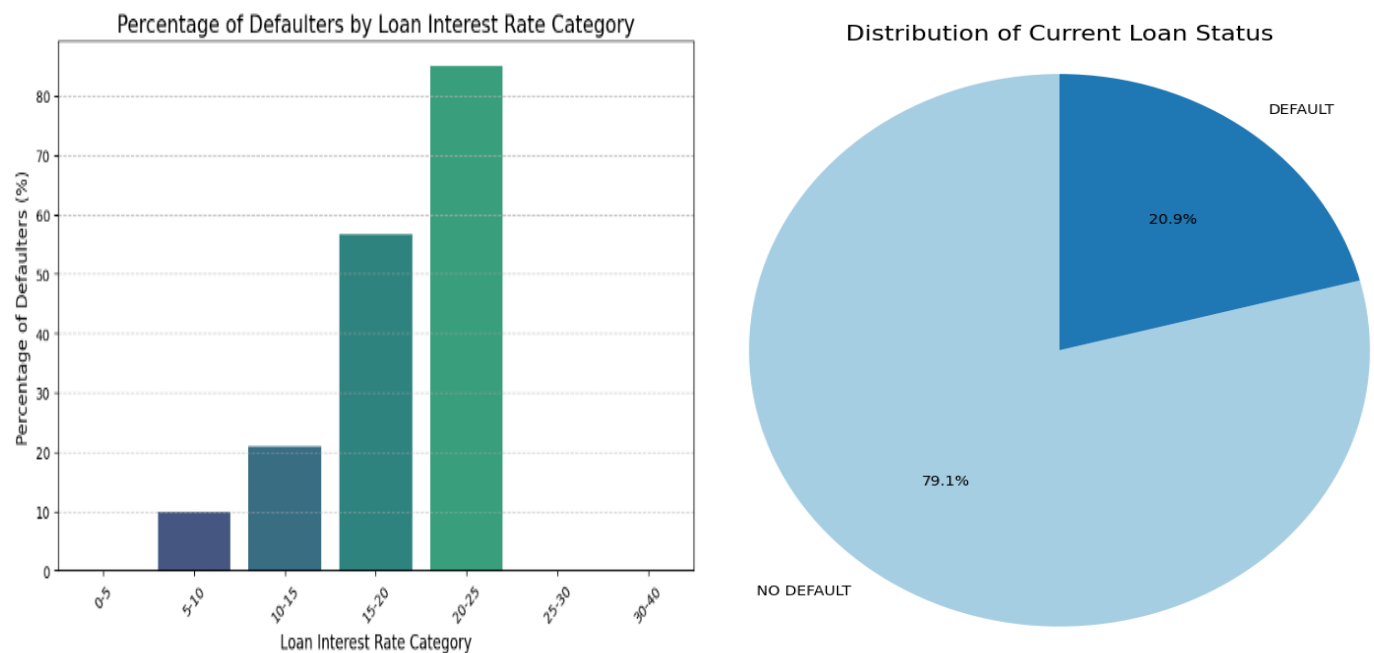


Figure 9 : A)Percentage of defaulters by Loan Interest Rate Category B)Distribution of Current Loan Status

The bargraph highlights a clear relationship between interest rates and loan default rates. As interest rates increase, the likelihood of loan defaults rises significantly. A notable proportion of defaults is observed in the 20-30% interest rate range, indicating that borrowers struggle more with higher interest burdens. In contrast, lower interest rates (0-10%) are associated with fewer defaults, suggesting greater affordability for borrowers. This trend underscores the strong correlation between higher interest rates and increased default risk.

This pie chart provides a visual representation of the distribution between default and non-default loan statuses, helping identify the proportion of customers with loan defaults in the data.

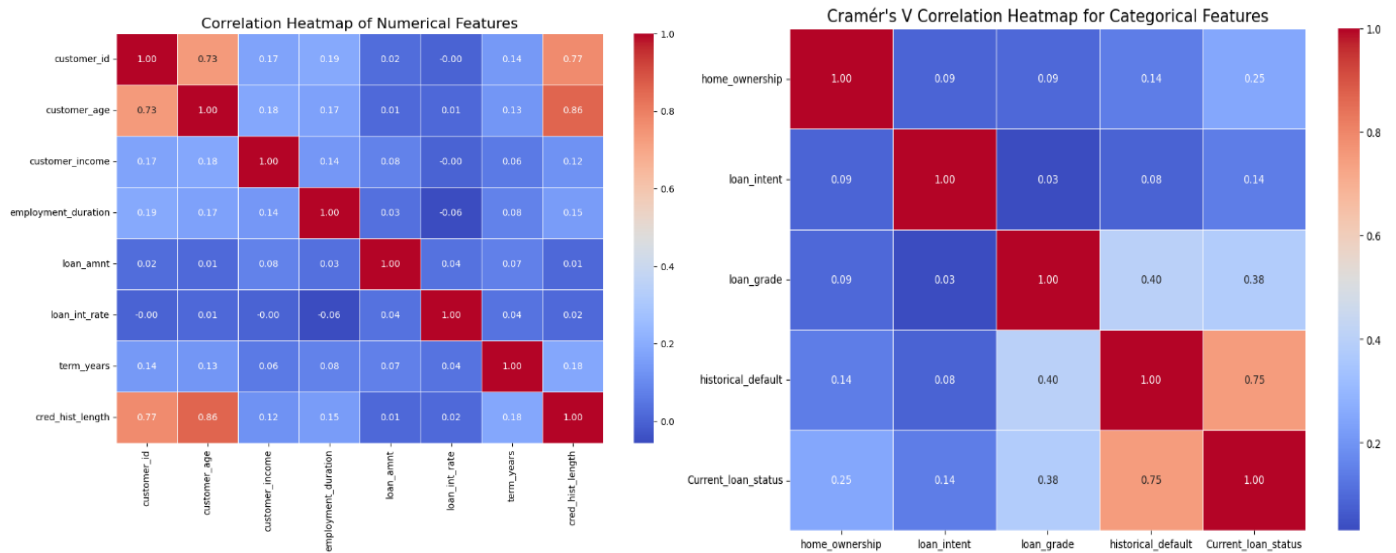


Figure 10: A)Correlation Heatmap of Numerical Features B)Correlation Heatmap of Categorical Features

The first correlation highlights significant correlations among features, with a strong positive relationship observed between customer age and credit history length (0.86) and between loan amount and customer income (0.77). These correlations suggest that older customers with longer credit histories and higher incomes are likely to take larger loans. In contrast, employment duration shows weak correlations with other features, indicating its limited influence on loan behavior. These insights aid in feature selection by identifying variables with substantial predictive power, enhancing model accuracy and efficiency.

The second correlation highlights important correlations among features. A strong correlation of 0.75 between historical default and current loan status suggests that past defaults have a significant impact on current loan outcomes. There is a moderate correlation between loan grade and both current loan status (0.38) and historical default (0.40), indicating that loan grade plays a role as a risk indicator. On the other hand, features like home ownership and loan intent show weak correlations, implying they have minimal impact on predicting loan defaults. These insights help prioritize the most relevant variables for predictive modeling, enhancing the model's efficiency.

6. Feature Engineering

Our team is using a combination of feature engineering and credit scoring techniques to enhance the performance and interpretability of our loan default prediction model. We start by transforming the target variable, *Current_loan_status*, into binary values to simplify the model: 'NO DEFAULT' becomes 0 and 'DEFAULT' becomes 1. For the *historical_default* column, which initially contains the values 'N', 'Y', and 'Unknown', we handle missing or unknown values carefully. We replace 'Unknown' with 'N' and temporarily change 'N' to 'temp' to avoid overlap during processing. Then, a balanced number of 'Y' and 'N' values are randomly assigned to the rows with the 'Unknown' category, ensuring data integrity.

We also transform categorical variables like *home_ownership*, *loan_intent*, and *loan_grade* into numerical representations using Label Encoding, making them more accessible to machine learning models. Additionally, we compute the **Credit Utilization Ratio** (loan amount divided by customer income), which isn't directly used in the model but plays a critical role in evaluating credit scores. Raw credit scores are normalized to a standard range of 300–850 for better interpretation, and loan grades like A, B, or F are converted into numerical scores (e.g., A → 100, F → 10) to quantify creditworthiness.

To gain deeper insights into a customer's financial behavior, we calculate their **credit score** using five key factors. Kabilan worked on the credit score calculation part, ensuring accuracy and alignment with industry standards.

6.1 Payment History (200 Points):

This considers whether the customer has defaulted before (*current_loan_status*) and their past defaults (*historical_default*). A perfect payment record earns the maximum points.

6.2 Credit Utilization (150 Points):

This measures how much credit a customer uses compared to their income.

Formula:

$$\text{Credit Utilization Ratio} = \text{customer income/loan amount}$$

$$\text{Credit Utilization Score} = 150 \times (1 - \min(\text{Credit Utilization Ratio}, 1))$$

Lower utilization gets a higher score, as it shows better credit management.

6.3 Credit History Length (100 Points):

This reflects how long a customer has been using credit.

Formula:

$$\text{Credit History Length Score} = \min(\text{cred_hist_len} \times 20, 100)$$

Longer histories get higher scores, up to 100 points.

6.4 Employment Stability (100 Points):

This measures job stability based on how long the customer has been employed.

Formula:

$$\text{Employment Stability Score} = \min(\text{employment_duration} / 12 \times 10, 100)$$

Longer employment earns more points.

6.5 Loan Grade and Interest (100 Points):

This scores the credit grade assigned to the loan.

Mapping:

- A → 100 points
- B → 80 points
- C → 60 points
- D → 40 points
- E → 20 points
- F → 10 points

The **final credit score** is calculated as:

$$\text{Credit Score} = X + (\text{Payment History Score} + \text{Credit Utilization Score} + \text{Credit History Length Score} + \text{Employment Stability Score} + \text{Loan Grade Score})$$

Here, XXX is the base score, ensuring the final value aligns with the standard range of 300–850.

Based on the score, we classify customers into risk categories:

- **Excellent (800+):** Very low risk
- **Very Good (740–799):** Low risk
- **Good (670–739):** Moderate risk
- **Fair (580–669):** High risk
- **Poor (below 580):** Very high risk

By using this score alongside the model's predictions, we can better assess the chances of a customer defaulting on their loan.

To refine the model's performance, we consider feature interactions, such as the relationship between credit utilization and employment stability, to better assess default risk. Customers with high credit utilization and low employment stability are weighted more heavily as higher risk. Through exploratory data analysis (EDA), we identify predictive and meaningful features, optimizing the model by adding or removing features based on performance. Additionally, integrating a credit score calculation alongside the machine learning model enhances interpretability, making the system more transparent for stakeholders. This approach not only improves prediction accuracy but also ensures compliance with regulatory standards, fostering trust and fairness in lending decisions.

The model evolves through regular validation to ensure robustness across customer segments and loan types. By combining machine learning with traditional credit scoring, it balances predictive power and transparency. This hybrid approach improves risk classification

and facilitates adoption by financial institutions, aligning with established practices while enhancing predictions.

7. Model Development and Training

After completing preprocessing, exploratory data analysis, and feature engineering, our team focused on selecting the most suitable machine learning model to achieve high accuracy for our dataset. Recognizing the importance of testing a variety of models, we divided the task of training and evaluating four different models among team members. Akash Narayan took responsibility for XGBoost, a powerful boosting algorithm known for its strong performance on structured data, particularly when that data includes categorical features. Jayaram focused on Logistic Regression, a straightforward and interpretable model often used for binary classification. HariKrishnan was assigned with Random Forest, an ensemble model that reduces overfitting, and Shikhar Pandey was assigned with Neural Networks, which can learn complex patterns in data.

7.1 Random Forest Model :

HariKrishnan worked on the Random Forest model, which combines multiple decision trees for better accuracy and reduced overfitting. He prepared the data by encoding categorical features into numbers and splitting it into training and testing sets. He trained the model using optimal settings like 100 trees and evaluated its performance with accuracy and a classification report. His work ensured a reliable and robust model that not only performed well but also provided insights into the importance of different features.

7.2 Logistic Regression:

Jayaram worked on the Logistic Regression model, a straightforward and easy-to-understand method for predicting binary outcomes. He prepared the data by converting categorical features into numbers and splitting it into training and testing sets. He trained the model with the right settings and checked its performance using accuracy and a classification

report. His work produced a reliable and easy-to-interpret model, showing how different features affect loan defaults.

7.3 Neural Networks :

Shikhar Pandey worked on the Neural Network model to find patterns in the data for predicting loan defaults. He prepared the data by converting categories to numbers, normalizing the values, and splitting it into training and testing sets. He built a network with two hidden layers, used the Adam optimizer, and trained the model for several rounds. His work created a strong model that could learn complex relationships in the data and give accurate predictions.

7.4 XGBoost :

Akash Narayan worked on training the XGBoost model . He handled preprocessing tasks such as encoding categorical variables and resampling the data using SMOTE to address class imbalance. After splitting the data into training and testing sets, Akash trained the XGBoost model, evaluated its performance, and generated the classification report and accuracy score.

Comparison and Analysis:

Model	Accuracy	Class 0 Precision	Class 0 Recall	Class 1 Precision	Class 1 Recall	F1-Score (Class 0)	F1-Score (Class 1)
Neural Network	95.74%	0.91	0.88	0.97	0.98	0.89	0.97
XGBoost	98.04%	0.98	0.98	0.98	0.98	0.98	0.98
Random Forest	~96.5%	~0.92	~0.88	~0.97	~0.98	~0.90	~0.97
Logistic Regression	~95.5%	~0.90	~0.86	~0.97	~0.98	~0.88	~0.97

Figure 10: Table with Comparison and Analysis of different classifier models

- XGBoost performs the best overall, with the highest accuracy (98.04 %) and F1-scores for both classes, especially excelling at predicting defaulters (Class 1). It has high precision and recall across the board, making it the most reliable for this task.
- Neural Network follows closely, with a slightly lower accuracy (95.74%) and good F1-scores, but its performance on non-defaulters (Class 0) is slightly weaker compared to XGBoost.
- Random Forest offers a similar performance to the Neural Network, with slightly lower overall accuracy but still achieving high precision and recall for defaulters.
- Logistic Regression has the lowest accuracy (~95.5%) and the weakest performance for non-defaulters, though it still provides decent results for defaulters.

XGBoost stands out as the best model due to its high accuracy and strong balance between precision, recall, and F1-score, especially for defaulters.

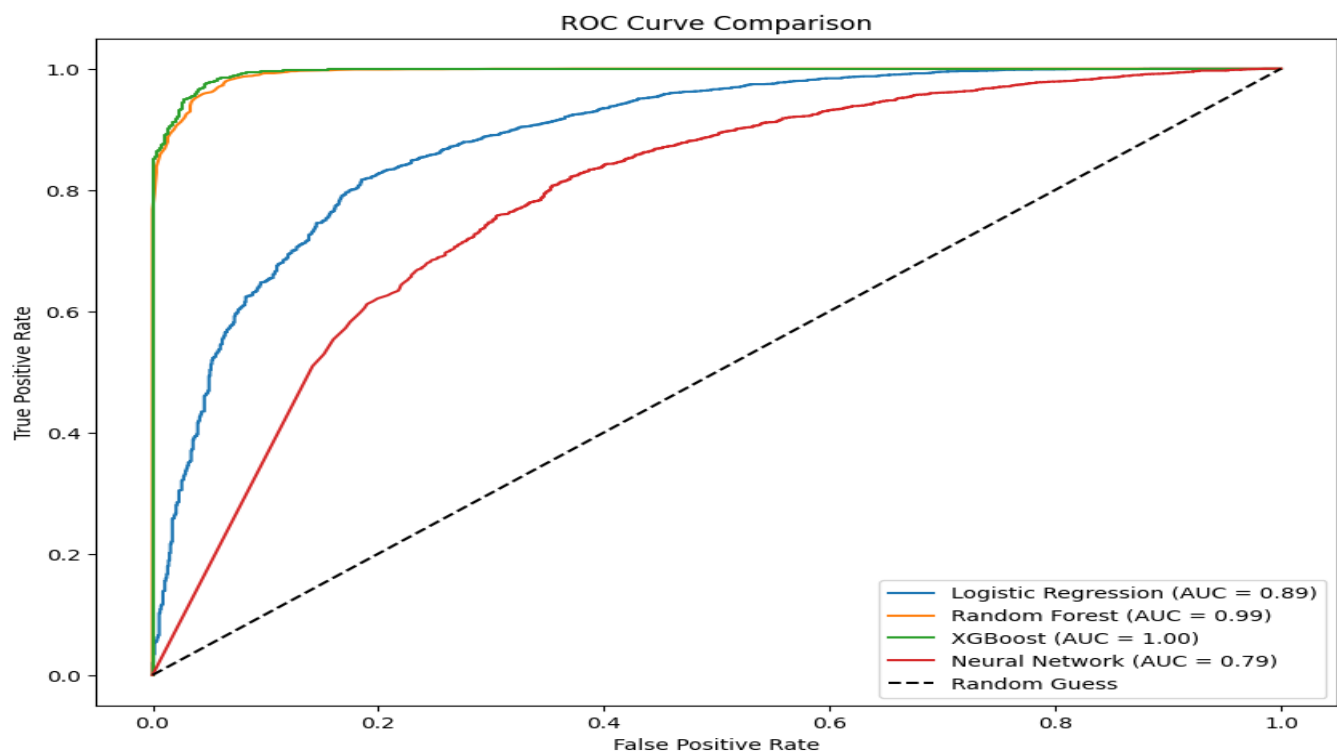


Figure 11: ROC Curve Comparison Across Classifier Models

The ROC (Receiver Operating Characteristic) curve plots the trade-off between the true positive rate (sensitivity) and false positive rate (1-specificity) for each model at various

thresholds. The AUC (Area Under the Curve) value represents the model's ability to distinguish between the positive and negative classes, with higher values indicating better performance.

8. Development and Functionality of the Application

Our team developed the Home Credit Default Risk application by combining machine learning models with a user-friendly web interface to deliver accurate predictions and insights. After preprocessing the dataset (HCD(Preprocessed) (1).csv), we trained and evaluated various models, including Logistic Regression, Random Forest, XGBoost, and Neural Networks, ultimately selecting XGBoost for its superior performance. The pre-trained model, stored in *cat_model.pkl*, and other essential files like *label_encoders.pkl* were integrated into the backend for consistent and efficient input handling.

The backend, implemented in *app.py* using Flask, manages the core logic of the application. It processes user inputs, validates and preprocesses the data, performs predictions using the trained model, and calculates credit scores based on features like payment history, loan amount, and credit history. These predictions and scores are then sent to the frontend for display.

The frontend, designed with *index.html* and styled using *style.css*, serves as the user interface. It allows users to input details such as loan amount, income, and employment duration. After submission, the backend processes these inputs and returns the results, which are displayed in an intuitive format. The results include loan default predictions, credit scores, and their corresponding categories, along with potential visualizations to enhance user understanding.

The application underwent iterative improvements, with updates reflected in files stored in the *.history* folder. This collaborative effort resulted in a fully functional web-based tool that provides users with actionable insights into their credit standing and default risk, ensuring both accuracy and ease of use.

9. Literature Survey

The Home Credit Default Risk Analysis project leverages machine learning to predict loan defaults by analyzing customer demographics, financial histories, and credit behaviors, addressing gaps in the financial sector and enabling informed lending decisions. It utilizes advanced algorithms like Random Forest, XGBoost, and Neural Networks, offering improved accuracy and insights over traditional methods like Logistic Regression. The project incorporates feature engineering, processing variables such as income, loan amounts, and credit history length, and creates derived metrics like credit utilization ratios for better modeling. Implemented as a secure Flask web application, it integrates a credit scoring mechanism assessing payment history and employment stability, ensuring transparency, trust, and regulatory compliance through encryption techniques.

The Home Credit Default Risk Analysis project demonstrates XGBoost's superior performance in predicting loan defaults, particularly with imbalanced datasets, as highlighted in the [Comparative Study of Credit Risk Evaluation for Unbalanced Datasets Using Deep Learning Classifiers](#) on IEEE Xplore. This study emphasizes XGBoost's gradient boosting technique for capturing complex patterns in financial data. Additionally, feature engineering plays a crucial role in improving model performance, contributing up to a 15% enhancement, supported by the [Application of Machine Learning in Credit Risk Assessment: A Prelude to Smart Banking](#), which underscores the value of derived features like credit utilization and employment stability in boosting predictive accuracy. These studies provide valuable insights into the techniques driving modern credit risk analysis systems.

The project emphasizes security and compliance with global standards, ensuring trust in financial transactions. Future enhancements include to [Explainable AI in Credit Risk](#) for interpretability and [Mobile Integration for Financial Systems](#) for broader accessibility, along

with usability upgrades for non-technical users and periodic model retraining to maintain relevance in dynamic markets

10. Output Screens

10.1 User Interface:

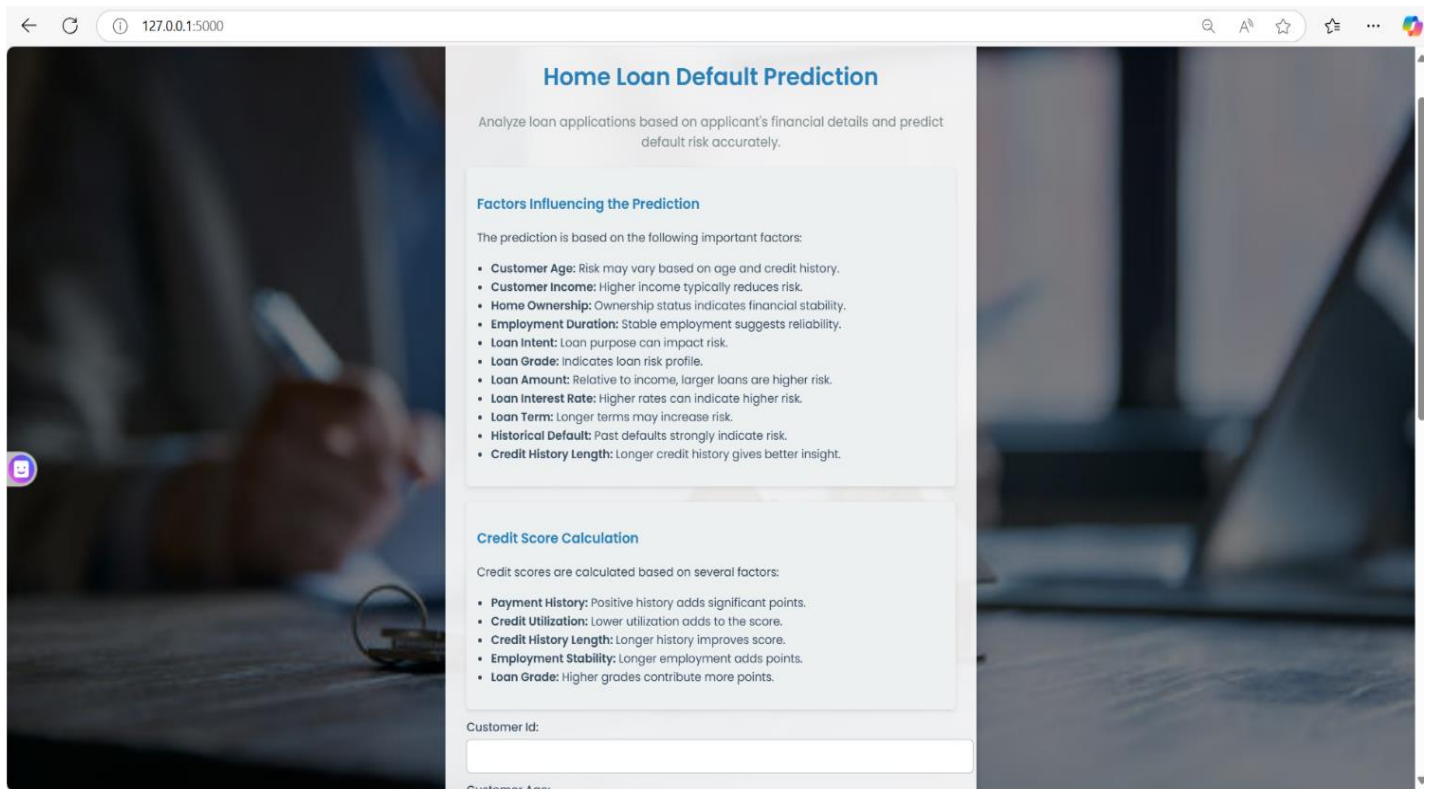
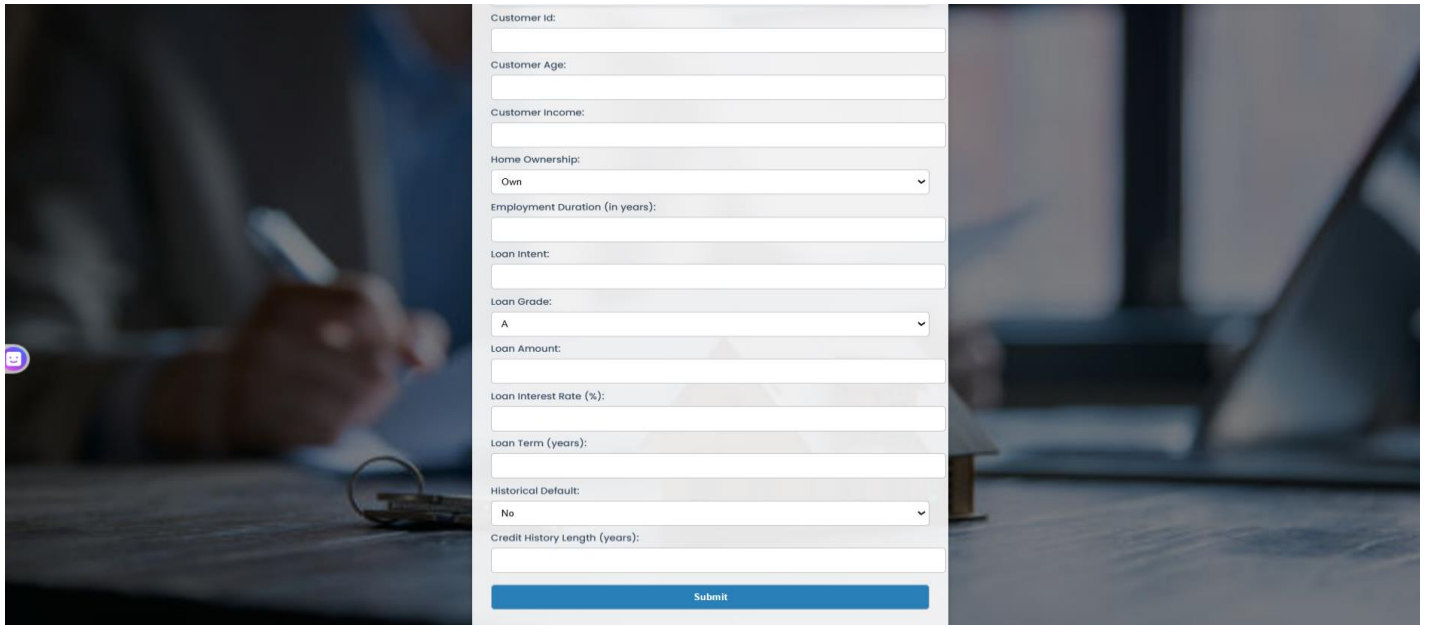


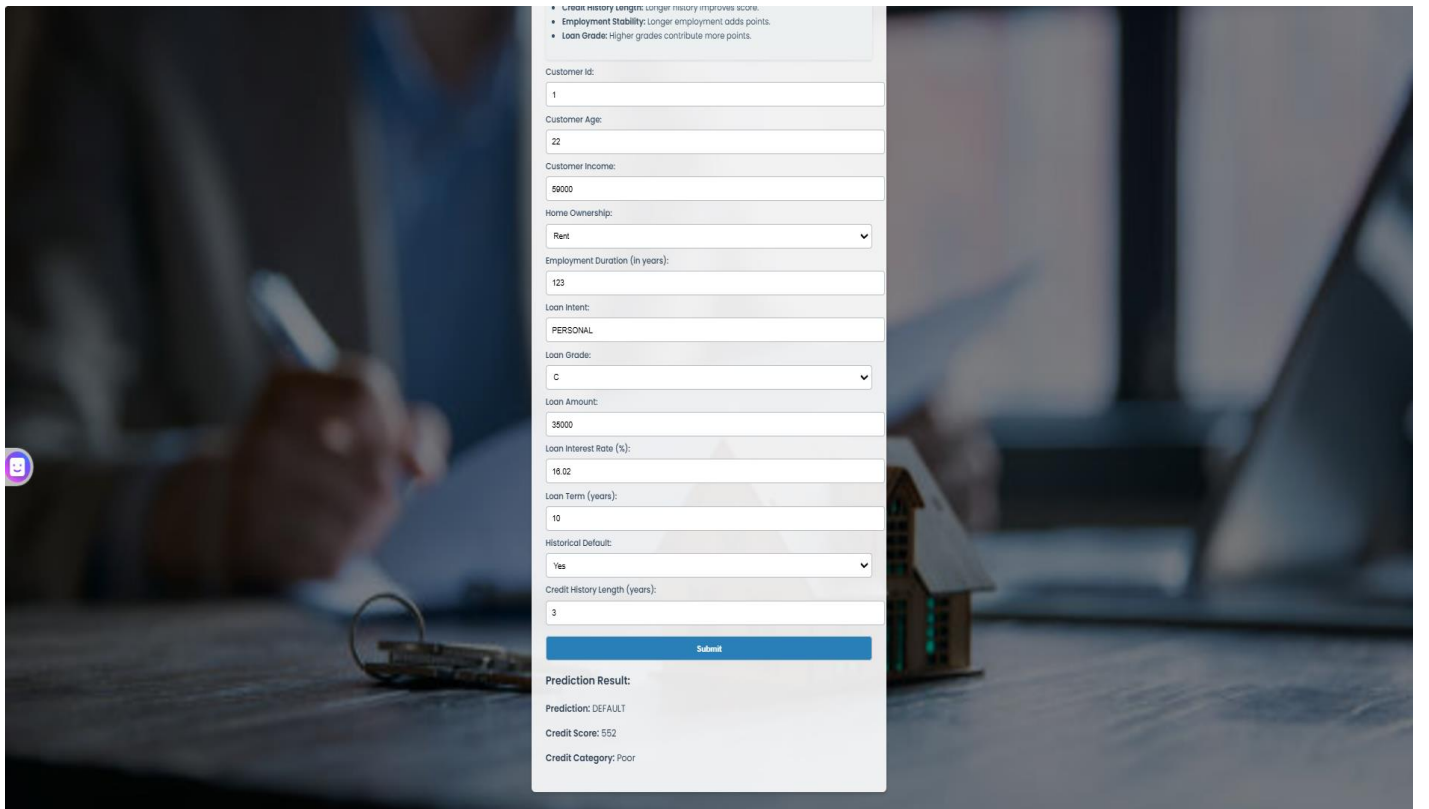
Figure 12 :Index page with description of features



The image shows a web form for loan application. The form is centered on a background image of a person writing on a document. The form fields are as follows:

- Customer id:
- Customer Age:
- Customer income:
- Home Ownership:
- Employment Duration (in years):
- Loan Intent:
- Loan Grade:
- Loan Amount:
- Loan Interest Rate (%):
- Loan Term (years):
- Historical Default:
- Credit History Length (years):
-

Figure 13 :Index page with input details



The image shows the same web form as Figure 13, but with the following input values filled in:

- Customer id: 1
- Customer Age: 22
- Customer income: 50000
- Home Ownership: Rent
- Employment Duration (in years): 123
- Loan Intent: PERSONAL
- Loan Grade: C
- Loan Amount: 35000
- Loan Interest Rate (%): 16.02
- Loan Term (years): 10
- Historical Default: Yes
- Credit History Length (years): 3
-

Below the form, the prediction results are displayed:

Prediction Result:
Prediction: DEFAULT
Credit Score: 552
Credit Category: Poor

Figure 14: Prediction Result after submitting input details

11. Conclusion

In conclusion, the Home Credit Default Risk Analysis project showed that machine learning models, especially XGBoost, are effective in predicting loan defaults. Feature engineering also played a key role, improving the model's performance by up to 15% through features like credit utilization and employment stability. The results highlight the importance of using strong algorithms and well-designed features to improve accuracy in credit risk prediction. Overall, the project provides valuable insights into modern methods for assessing credit risk, helping to create more accurate and efficient credit evaluation systems.

Additionally, the project emphasizes the significance of data preprocessing and feature selection in improving model performance. By handling missing data, encoding categorical variables, and normalizing numerical features, the models were able to learn better patterns from the dataset, leading to more reliable predictions. The process of feature selection, particularly in identifying the most influential variables, proved to be crucial in reducing overfitting and enhancing the model's generalization capabilities. These steps demonstrate how careful data preparation can significantly contribute to the success of machine learning applications in real-world scenarios like credit risk analysis.

Furthermore, the integration of a secure and user-friendly Flask web application allows for seamless interaction with the model, making the prediction process accessible to stakeholders in the financial sector. The inclusion of robust encryption ensures that sensitive customer data is handled with the utmost security, which is vital in maintaining trust and compliance with financial regulations. This project not only advances the use of machine learning in credit risk prediction but also showcases the practical application of modern technologies in the financial services industry, providing a foundation for future developments in the field.

12.Future Scope

The Home Credit Default Risk Analysis project holds significant potential for future advancements and applications, particularly in the rapidly evolving domain of financial technology. Here are some key areas of its future scope:

1. Enhanced Risk Prediction Models:

With the continuous evolution of machine learning and AI, more advanced algorithms like deep learning and ensemble methods can be utilized to further improve the accuracy of credit risk predictions. This could lead to better identification of high-risk borrowers and minimize default rates.

2. Real-Time Risk Assessment

By integrating this system into live banking and credit systems, lenders can perform real-time risk assessments during the loan application process. This would speed up decision-making and reduce manual intervention.

3. Integration with Financial Ecosystems

The system can be integrated with other financial platforms, such as payment gateways and insurance services, to provide a holistic view of a customer's financial behavior. This interconnected data can improve the robustness of risk assessment.

4. Global Adaptability

By customizing models for different regions and demographics, the system can be extended to assess credit risk in emerging markets where traditional credit scoring systems are ineffective.

5. AI-Powered Insights for Financial Education

Borrowers can benefit from insights provided by AI models, helping them understand factors contributing to their credit scores. This can promote financial literacy and encourage responsible borrowing.