Atal Bihari Vajpayee Indian Institute of Information Technology and Management, Gwalior

Internet Technology for Business
Report

Topic:- Home Loan Approval Prediction System
Group no. :- 4

Submitted To:- Dr. Vinay Singh

Submitted By:- Gourav Waskle (2020IMG-024)
Sourabh Singh Jindal (2020IMG-062)

Table of Content

Introduction

In the realm of financial necessities, loans serve as indispensable tools for diverse purposes, ranging from education to the acquisition of homes or vehicles. However, the evaluation of loan eligibility constitutes a multifaceted task. This project, titled "Home Loan Approval Prediction," harnesses the power of Machine Learning and Python to streamline and simplify this intricate process. The primary focus is on predicting whether an applicant's profile aligns with the criteria for loan approval.

Traditionally, such decisions have been entrusted to bank personnel, a process that can be both sluggish and susceptible to biases. The overarching objective of this endeavor is to revolutionize the loan approval landscape. By leveraging advanced technologies, the project aims to enhance the efficiency and accuracy of the approval process, reducing the reliance on human judgment and mitigating potential biases. The ultimate vision is to foster the creation of a financial system that is not only more efficient but also fair and accessible to all, thereby contributing to the establishment of a more equitable and supportive economic ecosystem.

Motivation and Problem Statement

The motivation for this project has been derived from various factors. Some are mentioned below:-

1. **Enhance efficiency :** Automating the loan approval process saves time and effort, benefiting both bank staff and customers.
2. **Speed up service :** Automation can significantly improve the speed of service provided to customers, reducing processing time.
3. **Improve accuracy :** Machine learning algorithms can provide more accurate predictions for loan approval based on applicant information.
4. **Complex decision-making :** Loan approval involves evaluating various aspects of an applicant's profile, making it a challenging process for banks.
5. **Digital transformation :** The financial sector is undergoing a digital transformation, with AI and ML leading this revolution.
6. **Subjectivity in the old process :** The old loan approval process relies on subjective human judgement, leading to potential unfairness and inconsistency.
7. **Create a transparent system :** The project aims to establish a transparent, efficient, and inclusive financial system, benefiting both lenders and borrowers.

The old system of loan approval involves bank managers, customers, a lot of paperwork, time, money and energy which can be saved by the bank by using a machine learning model which can help in making a decision. This reduces the overload of managers and also eases

the process for the customer and also makes it cost effective, efficient for the Bank. The problem statement for a loan approval prediction system typically involves designing a model that can assess whether a loan application should be approved or denied based on various factors.

The main objective of this project is to develop a robust and accurate loan approval prediction system. The system should analyse loan applications and provide a prediction on whether the application should be approved or denied, taking into consideration various factors and risk indicators. The expected outcome is a reliable loan approval prediction system that can assist financial institutions in making informed decisions, improving efficiency, and reducing the risk of default.

About Dataset

The dataset contains the information of 1033 loan applications which were either accepted or rejected. The dataset contains the following key features:

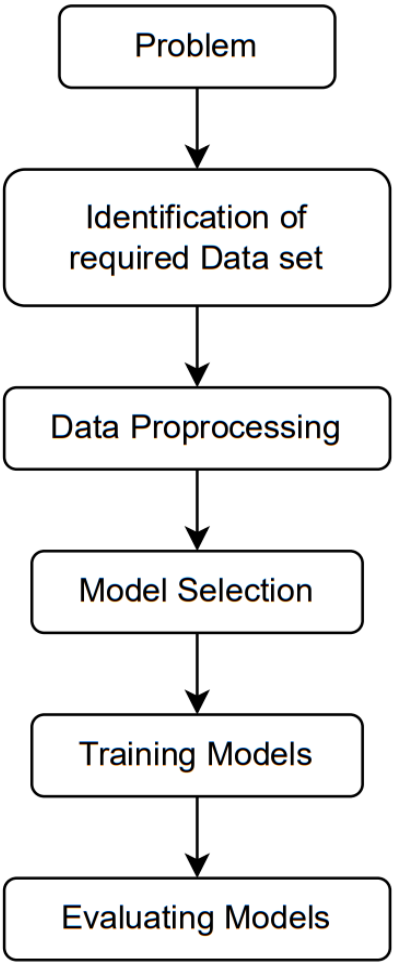
1. Loan_id : A unique id given to each customer.
2. Gender : Gender of the applicant Male/female.
3. Married : Marital Status of the applicant, values will be Yes/ No
4. Dependents : It tells whether the applicant has any dependents or not.
5. Education : It will tell us whether the applicant is Graduated or not.
6. Self_Employed : This defines that the applicant is self-employed i.e. Yes/ No
7. ApplicantIncome : Applicant income
8. CoapplicantIncome : Co-applicant income
9. LoanAmount : Loan amount (in thousands)
10. Loan_Amount_Term : Terms of loan (in months)
11. Credit_History : Credit history of individual’s repayment of their debts.
12. Property_Area : Area of property i.e. Rural/Urban/Semi-urban

The dataset is split into two subsets: the training set and the testing set. This is typically done to assess how well a machine learning model can generalise from the data it has seen during training to new, unseen data. In this case, 70% of the data (approximately 723 entries) is used for training the models, and the remaining 30% (approximately 310 entries) is used for testing.

Proposed Methodology

Frame Work

Training a machine learning model for binary classification involves several key processes. These processes collectively transform raw data into a predictive model capable of classifying data points into one of two distinct classes. Given below are the essential steps:



Data Collection and Preprocessing: The process begins with gathering and cleaning the dataset. Data preprocessing involves handling missing values, outlier detection, and transforming data into a suitable format. This step ensures that the data is ready for training and is free from noise or inconsistencies that could affect the model's performance.

Data Splitting: The dataset is divided into two parts: a training set (70%) and a testing set (30%). The training set is used to teach the model, while the testing set serves as a completely separate dataset for evaluating the model's performance.

Model Selection: Here we have used Logistic Regression and Random Forest Classifier. On small dataset Logistic Regression may work better then Random Forest Classifier and on large datasets vice-versa.

Training the Model: The selected model is trained on the training dataset, and this process involves finding the optimal parameters that minimise a predefined loss function. For example, in logistic regression, this entails finding the best weights for the features. In more complex models like Random Forest, optimization techniques such as gradient descent are employed.

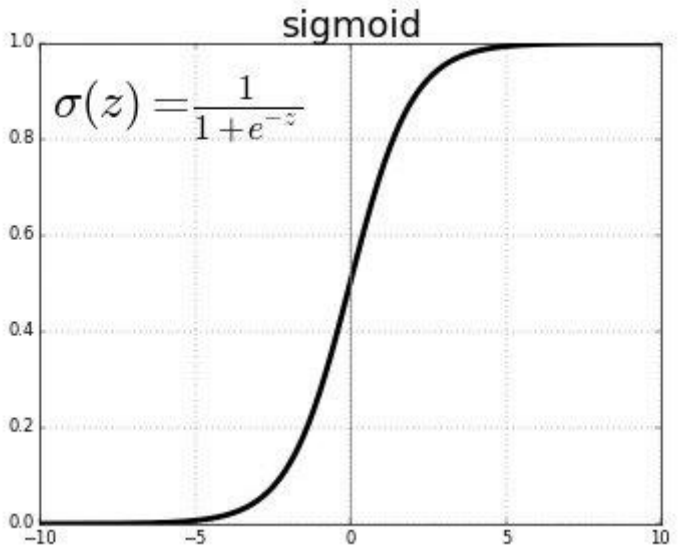
Model Evaluation: After training, the model's performance is assessed using the testing dataset. Common evaluation metrics for binary classification include accuracy, precision, recall. The choice of metric depends on the problem's goals and the balance between false positives and false negatives.

1. Logistic Regression:

Logistic regression is a statistical method used for analyzing a dataset in which there are one or more independent variables that can be used to predict the outcome of a categorical dependent variable. It is a fundamental and widely-used technique in the field of machine learning, statistics, and data science. Logistic regression is particularly suited for binary classification problems, here the dependent "Loan_Status" variable has two possible values, typically represented as 0 and 1, or "not approved" and "approved."

Basic Concept:
Logistic regression is an extension of linear regression, which is used for predicting continuous numeric values. In contrast, logistic regression predicts the probability that a given input belongs to a particular category or class.

Sigmoid Function:
At the core of logistic regression is the sigmoid function (also known as the logistic function). The sigmoid function takes any real-valued number and maps it to a value between 0 and 1. It has an S-shaped curve and is defined as follows:



In logistic regression, the hypothesis function *h(x)* is defined as the sigmoid function applied to the linear combination of the input features and their associated weights. Here x_1, x_2, x_3, …..x_n are the independent variables, and $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$,…….. $\boldsymbol{\beta}_n$ are the coefficients (weights) to be learned from the training data.

Cost Function:
The goal of logistic regression is to find the optimal values for the coefficients β that minimize the cost function. The cost function for logistic regression is known as the cross-entropy or log-loss function. It measures the difference between the predicted probabilities and the actual binary outcomes.

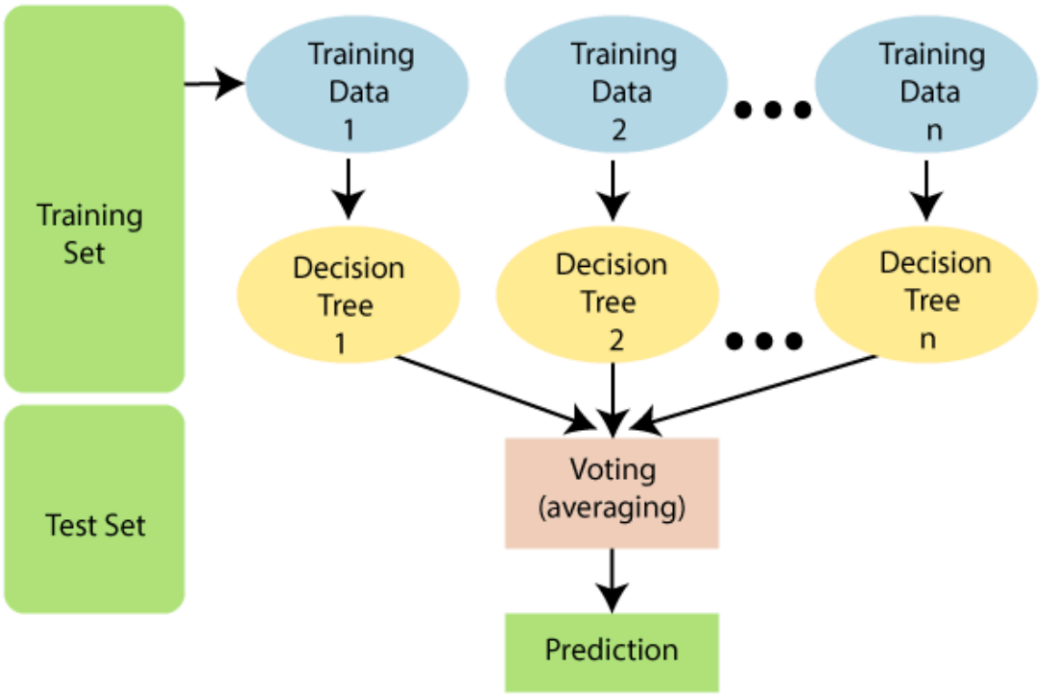
Optimization:
To find the optimal values for the coefficients that minimize the cost function, various optimization techniques can be used, such as gradient descent. Gradient descent iteratively updates the coefficients by taking small steps in the direction of steepest descent of the cost function.

Decision Boundary:
Once the coefficients are learned, the logistic regression model can be used to make predictions. The decision boundary is a threshold probability value (typically 0.5) that separates the two classes. If $h(x) \geq 0.5$, the data point is classified as 1; otherwise, it's classified as 0.

2. Random Forest Classifier:

A Random Forest classifier is an ensemble learning method based on decision tree models. It is a powerful and versatile machine learning algorithm that is widely used for both classification and regression tasks. Random Forest is known for its ability to produce accurate and robust predictions while mitigating some of the overfitting issues associated with individual decision trees. Below is the detailed explanation of the Random Forest classifier:

Ensemble Learning:
Random Forest belongs to the family of ensemble learning methods. Ensemble methods combine the predictions of multiple individual models to make more accurate and stable predictions. In the case of Random Forest, these individual models are decision trees.



Decision Trees:
Decision trees are a simple yet effective machine learning model that recursively splits the dataset into subsets based on the values of input features to make predictions. They work by partitioning the feature space into regions, with each region corresponding to a particular class or output value.

Randomization:
The key idea behind Random Forest is the introduction of randomness in two main aspects:

- Random Subspace Selection: When creating each individual decision tree, only a random subset of the available features is considered at each split. This helps decorrelate the trees and reduces overfitting.
- Bootstrap Aggregating (Bagging): Each tree is trained on a bootstrapped (randomly sampled with replacement) subset of the original training data. This results in variations in the training data for each tree.

Voting Scheme:
Random Forest combines the predictions of multiple decision trees by using a majority voting scheme in classification tasks. Each tree "votes" for a class, and the class with the most votes is considered the final prediction. In regression tasks, the predictions of individual trees are averaged to produce the final prediction.

Feature Importance:
Random Forest can also provide a measure of feature importance. It evaluates the contribution of each feature in making accurate predictions. Features that lead to the greatest reduction in impurity or the best splits in decision trees are considered more important.

Result

Dataset: The dataset consists of 1033 entries. Each entry likely represents a data point with various features or attributes, and these entries are labeled or categorized in some way.

Training and Testing Split: The dataset is split into two subsets: the training set and the testing set. This is typically done to assess how well a machine learning model can generalize from the data it has seen during training to new, unseen data. In this case, 70% of the data (approximately 723 entries) is used for training the models, and the remaining 30% (approximately 310 entries) is used for testing.

Models Used: Two different machine learning models have been employed for this task:

a. **Logistic Regression**: This is a statistical model used for binary classification problems. It calculates the probability of a data point belonging to a particular class and assigns it to the class with the highest probability. The model achieved an accuracy score of 80.13% on the test data. This means that, when the logistic regression model was evaluated on the 30% of data it hadn't seen during training, it correctly predicted the category or class for approximately 80.13% of the test data.

b. **Random Forest Classifier**: Random Forest is an ensemble learning technique that combines multiple decision trees to make predictions. It's often used for both classification and regression tasks. The Random Forest Classifier achieved an accuracy score of 76.01% on the test data. This means that, when this model was evaluated on the same test data, it correctly predicted the category or class for approximately 76.01% of the test data.

Interpreting the Accuracy Scores: The accuracy score is a common evaluation metric for classification models. It represents the proportion of correctly classified instances in the test set. In this case, the Logistic Regression model outperformed the Random Forest Classifier in terms of accuracy, with 80.13% accuracy compared to 76.01%. This suggests that, based on the available features and the nature of the problem, the Logistic Regression model may be a better choice for this specific task.

Confusion Matrix:

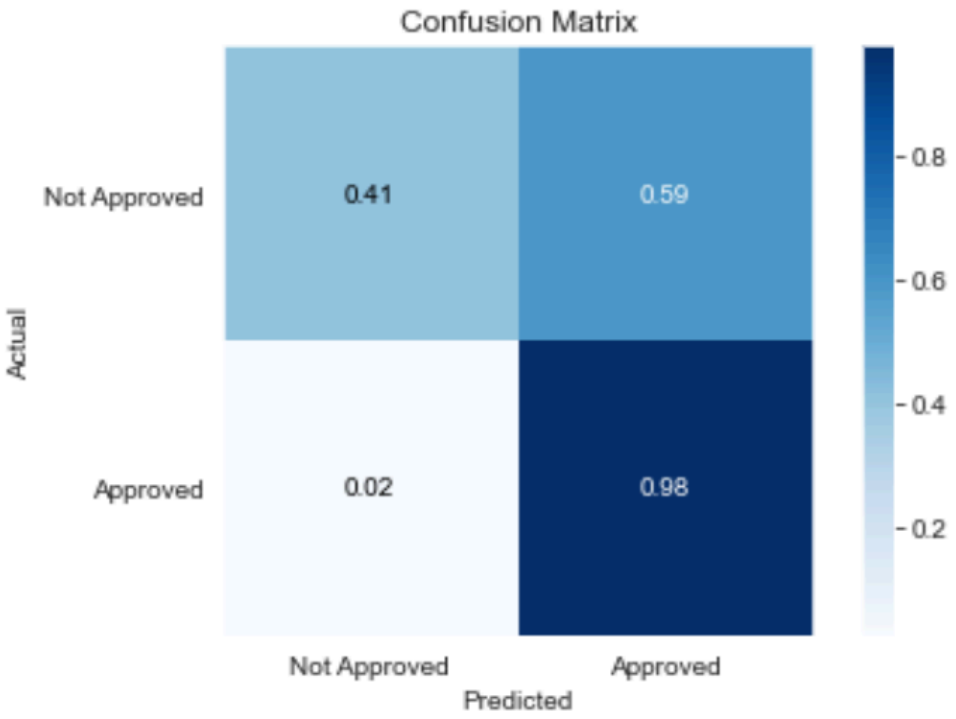
A confusion matrix, also known as an error matrix, is a performance evaluation tool used in machine learning and classification tasks to assess the performance of a classification model. It provides a detailed summary of how well a classification model has performed by comparing its predictions to the actual ground truth values. The confusion matrix is especially useful when dealing with binary classification problems.

A confusion matrix is typically presented in a table format with four main components:

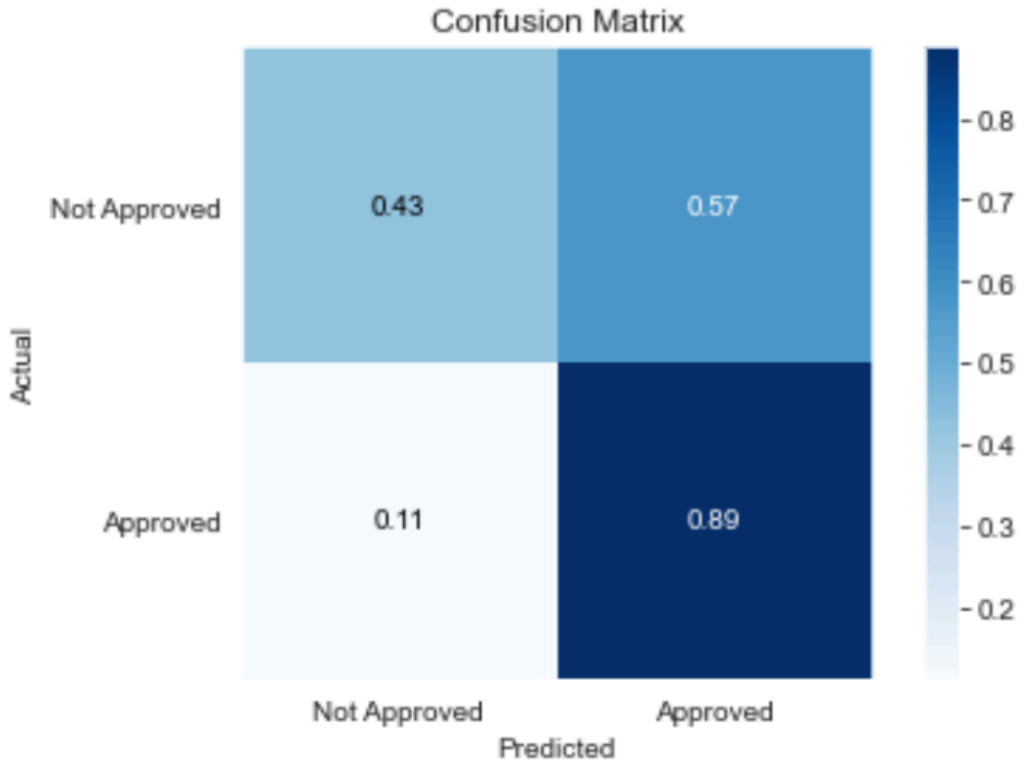
1. **True Positives (TP)**: These are the cases where the model correctly predicted the positive class. In other words, the model predicted a positive outcome, and the actual outcome was also positive.
2. **True Negatives (TN)**: These are the cases where the model correctly predicted the negative class. The model predicted a negative outcome, and the actual outcome was also negative.
3. **False Positives (FP)**: These are the cases where the model incorrectly predicted the positive class when it should have predicted the negative class. This is also known as a Type I error or a "false alarm."
4. **False Negatives (FN)**: These are the cases where the model incorrectly predicted the negative class when it should have predicted the positive class. This is also known as a Type II error or a "miss."

So for the two model Confusion matrix is given below:

1) Logistic Regression:



2) Random Forest Classifier:



Conclusion

- By implementing a predictive model for loan approval, financial institutions and banks can automate and streamline their loan approval processes. This automation can significantly reduce the time and resources required for manual review and decision-making.
- Automation also reduces the likelihood of human error in the decision-making process, leading to more consistent and reliable loan approval decisions.
- The predictive model acts as a valuable tool for managing risk. It uses data-driven insights to assess the creditworthiness of loan applicants. This means that institutions can make more informed and objective decisions about whether to approve a loan.
- By reducing the chances of approving loans to individuals or businesses with a higher risk of default, financial institutions can minimize their exposure to non-performing loans. This, in turn, helps to protect their financial stability.
- Implementing a predictive model that accurately identifies high-risk applicants can lead to a reduction in loan default rates. This is a critical benefit for financial institutions, as lower default rates contribute to improved profitability and sustainability.
- Fewer defaults mean that financial institutions have to allocate fewer resources for debt collection efforts, legal proceedings, or write-offs, ultimately saving time and money.
- The use of a predictive model enables financial institutions to provide loan approval decisions much faster than traditional manual processes. Applicants can receive quick responses, enhancing their experience.
- With data-driven decision-making, financial institutions can provide applicants with transparent explanations for loan approval or rejection. This transparency builds trust and credibility with customers, as they can better understand the basis of the decision.
- Efficiency in decision-making, risk assessment, and reduced defaults can lead to significant cost savings for financial institutions. They can allocate resources more efficiently, potentially reducing the need for extensive manual underwriting or expensive loan approval processes.
- The use of predictive models in loan approval can assist financial institutions in complying with regulatory requirements. These models can incorporate regulatory guidelines and ensure that lending decisions align with legal and ethical standards.

Novelty and Future Scope

The integration of cutting-edge technologies and innovative strategies can propel the Housing Finance Companies loan eligibility prediction model to new heights. By envisioning and implementing future advancements, the company can stay ahead of the curve and revolutionise the home loan application process.

1. Integration with Financial Data:

In the future, the Housing Finance Company could elevate its loan eligibility prediction model by integrating real-time financial data from diverse sources. For instance, consider linking the system to dynamic data such as stock market indices, currency exchange rates, and individual investment portfolios. This integration would provide a more comprehensive and up-to-the-minute view of a customer's financial health. By considering the latest financial information, the model ensures a more accurate assessment of loan eligibility.

For example, if a customer's investment portfolio experiences a sudden surge, reflecting improved financial stability, the real-time integration would capture this positive change, potentially impacting the loan eligibility decision in their favor. This integration not only enhances the precision of the model but also aligns the company with the rapidly evolving landscape of financial technology.

2. Real-Time Processing:

The future of the loan eligibility prediction system involves a paradigm shift towards real-time processing capabilities. Imagine a scenario where, upon submitting a home loan application, customers receive an instantaneous eligibility decision. This not only reduces the processing time but also caters to the growing demand for quick and efficient services.

Consider a prospective homebuyer eager to secure a property. With real-time processing, they can swiftly receive feedback on their eligibility, allowing them to make informed decisions promptly. This acceleration in the application process not only enhances customer satisfaction but also positions the Housing Finance Company as a frontrunner in providing agile and responsive financial solutions.

3. Mobile Application Integration:

The future envisions a seamless and user-friendly loan application experience through mobile applications. Developing a mobile interface for the loan eligibility prediction system allows customers to conveniently input their details and receive instant feedback on their eligibility status. This not only aligns with the preferences of a digitally-driven consumer base but also expands the company's accessibility.

For instance, a customer commuting to work can initiate the loan application process through a mobile app during their commute, ensuring a more flexible and convenient user experience. The application could provide step-by-step guidance, simplifying the complex financial information input process. This level of accessibility and user-friendliness can significantly enhance customer engagement and contribute to a positive brand image.

4. Enhanced Risk Assessment:

Future advancements in the loan eligibility prediction model involve embracing sophisticated risk assessment techniques. The model could evolve to consider a broader range of factors, including dynamic risk profiles, market trends, and economic indicators. By incorporating machine learning algorithms that adapt to changing economic conditions, the accuracy of predicting loan eligibility can be substantially improved.

For example, during periods of economic uncertainty, the model could dynamically adjust risk thresholds, ensuring a more conservative approach to loan approvals. On the other hand, during economic upswings, the model could adopt a more lenient stance, facilitating increased approvals. This adaptability enhances the model's resilience to market fluctuations, providing a more robust and reliable risk assessment.

5. Personalized Loan Offers:

The future entails a shift towards personalised financial solutions, and the loan eligibility prediction model can lead this transformation. Imagine a scenario where the system goes beyond generic eligibility decisions and tailors loan offers based on individual customer profiles. By analysing a customer's financial history, behaviour, and preferences, the system could propose personalised loan terms, interest rates, and repayment plans.

For instance, a customer with a consistent repayment history and stable financial standing might receive a personalised offer with lower interest rates or extended repayment terms. This level of customization not only enhances customer satisfaction but also increases the likelihood of acceptance. Personalised loan offers reflect a customer-centric approach, acknowledging the unique financial situations and aspirations of each applicant.

6. Customer Feedback Loop:

Establishing a robust customer feedback loop is integral to the continuous improvement of the loan eligibility prediction system. This involves actively seeking and incorporating feedback from applicants regarding their experience with the loan application process. By understanding user perspectives, pain points, and preferences, the company can make informed refinements to the model and the overall application journey.

For example, if customers consistently provide feedback about a specific aspect of the application process being confusing, the company can implement user interface enhancements to address this concern. Additionally, gathering feedback on the accuracy of

predictions allows the company to fine-tune the model, ensuring it aligns more closely with the expectations and needs of the applicants. This iterative feedback loop creates a dynamic and responsive system that evolves in tandem with customer expectations.