

# Machine Learning-Based Loan Eligibility Prediction for Diverse Loan Types

## 1.Introduction:

Loans come in various forms, with each type requiring unique approval criteria based on diverse factors. This project uses machine learning to predict eligibility across four distinct loan types, analyzing custom features within each dataset to capture the nuances of each loan category. By understanding the specific patterns and parameters associated with different loans, our model delivers accurate, tailored predictions that improve decision-making. This approach not only enhances the precision of eligibility assessments but also streamlines the loan evaluation process, making it more adaptable and efficient for varied lending needs. With the use of advanced algorithms, the project offers a comprehensive solution for loan providers, enabling faster, data-driven decisions that better align with customer profiles. The model's ability to handle different loan types ensures it can cater to a wide range of financial products, promoting both fairness and transparency in the lending process. This makes the loan approval process more accessible, helping both institutions and borrowers make informed choices. The project focuses on four loan types: **Home Loan, Vehicle Loan, Personal Loan, and Other Loan.**

## 2. Project Overview

This project addresses loan eligibility prediction across four loan types: **Home Loan,**

### **Vehicle Loan, Personal Loan, and Other Loan.**

By analyzing customized feature sets for each loan category, this model aims to improve the accuracy and fairness of the loan evaluation process.

## 3. Data Preprocessing

Each loan type dataset is preprocessed independently to handle category-specific features, with steps including:

- **Data Cleaning:** Removal of null values, outliers, and duplicates.
- **Feature Engineering:** Customized features for each loan type to capture unique eligibility criteria.
- **Encoding and Scaling:** Encoding categorical features and scaling numerical features to enhance model performance.

## 4.Models Implemented

For each loan type, we have mainly implemented and compared the performance of:

1. **Logistic Regression**
2. **Random Forest Classifier**

These models were selected for their suitability for binary classification tasks, where we are predicting loan eligibility.

## 5. Performance Metrics

To evaluate model performance, we used the following metrics for each loan type:

- **Accuracy:** Measures the percentage of correct predictions.
- **Precision:** Evaluates how many of the positive predictions are correct.
- **Recall:** Measures the model's ability to identify all positive instances.

## 6. Data Splitting, Validation, Hyperparameter Tuning, and Overfitting Check

- **Data Splitting:** Split the dataset into **80%** training and **20%** testing using `train_test_split` and also splitting training set into **80%** training and **20%** validation
- **Hyperparameter Tuning:** Use **Grid Search** for optimal hyperparameters for **Random Forest** and calculating best mse for Logistic Regression
- **Overfitting Check:** Evaluate training vs testing accuracy to detect overfitting. A large difference between these indicates overfitting.

## Performance Results by Loan Type:

### 1. Loan Type: Home Loan

Sample Dataset:

Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area
Male	No	0	Graduate	No	5849	0		360	1	Urban
Male	Yes	1	Graduate	No	4583	1508	128	360	1	Rural
Male	Yes	0	Graduate	Yes	3000	0	66	360	1	Urban
Male	Yes	0	Not Graduate	No	2583	2358	120	360	1	Urban
Male	No	0	Graduate	No	6000	0	141	360	1	Urban

### Features and Label

- **Features:** The dataset includes the following features that influence home loan eligibility:
  - Gender, Married, Dependents, Education, Self\_Employed
  - ApplicantIncome, CoapplicantIncome, LoanAmount, Loan\_Amount\_Term
  - Credit\_History, Property\_Area
- **Label:** Loan\_Status is the target variable, indicating eligibility for a home loan (Y for eligible, N for ineligible).

### Models Used

1. **Batch Logistic Regression:** Employed for its simplicity and ability to interpret the influence of each feature on loan eligibility.

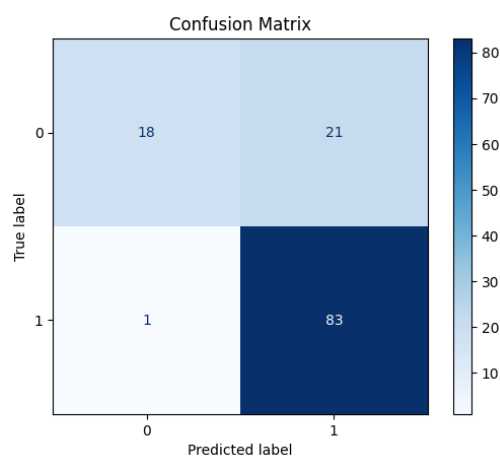
2. **Random Forest Classifier:** Selected to capture complex interactions among features and provide insights into feature importance.

### Performance Metrics

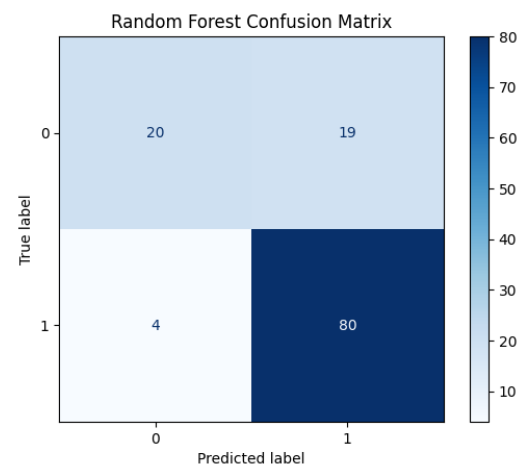
- **Batch Logistic Regression:**
  - Accuracy: 0.821
  - Recall: 0.988
- **Random Forest Classifier:**
  - Test Accuracy: 0.813
  - Recall: 0.952

### Confusion Matrices:

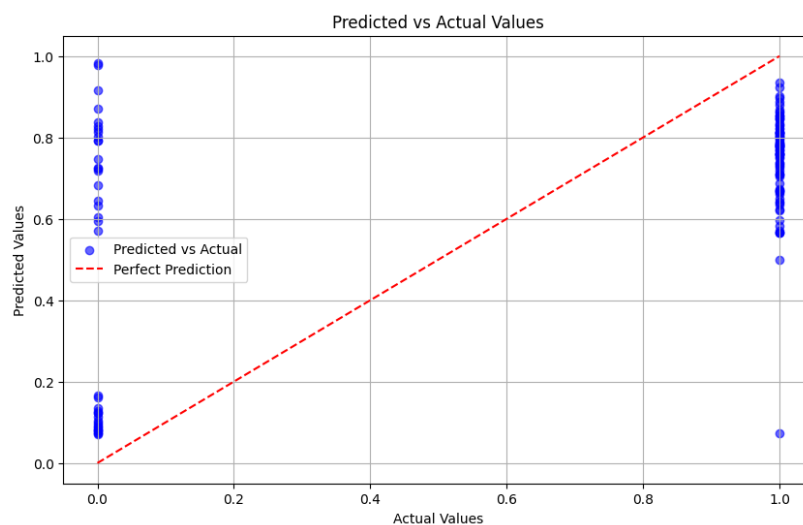
#### Logistic Regression:



#### Random Forest:



### Predicted vs Actual values:



**Observation:** Logistic Regression outperformed the Random Forest model in both accuracy and recall, indicating that a linear approach might be more suitable for this dataset.

## 2. Loan Type: Personal Loan

### Sample Dataset:

Age	Experience	Income	Family	CCAvg	Education	Mortgage	Securities Account	CD Account	Online
25	1	49	4	1.6	1	0	1	0	0
45	19	34	3	1.5	1	0	1	0	0
39	15	11	1	1	1	0	0	0	0
35	9	100	1	2.7	2	0	0	0	0
35	8	45	4	1	2	0	0	0	0

### Features and Label

- **Features:** The dataset includes the following features that influence personal loan eligibility:
  - Age, Experience, Income, Family, CCAvg (Credit Card Average Spending)
  - Education, Mortgage, Securities Account, CD Account, Online, CreditCard
- **Label:** Personal Loan is the target variable, where 1 represents an eligible applicant and 0 represents an ineligible applicant.

### Models Used

1. **Batch Logistic Regression:** Chosen for its simplicity and ability to provide probabilistic interpretations of the relationships between features and loan eligibility.
2. **Random Forest Classifier:** Used to capture complex relationships and interactions between the features, as well as for better generalization on unseen data.

### Performance Metrics

- **Logistic Regression:**
  - Accuracy: 0.927
  - Recall: 0.361
- **Random Forest Classifier:**
  - Test Accuracy: 0.987
  - Recall: 0.880

**Observation:** Random Forest significantly outperformed Logistic Regression in both accuracy and recall, indicating that the Random Forest model is better suited for this dataset, particularly for identifying personal loan applicants.

### Observations

1. **Model Performance:**
  - Random Forest achieved both a higher accuracy (0.987) and recall (0.880) compared to Logistic Regression.

- Logistic Regression had lower recall (0.361), which means it failed to identify a significant number of actual personal loan applicants. This suggests Random Forest's more complex decision-making process is better suited for this task.

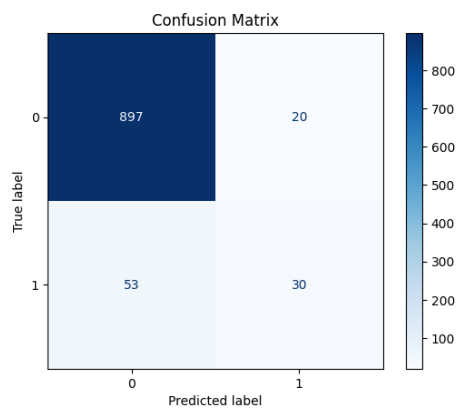
## 2. Feature Importance:

- Random Forest can provide insights into which features are most influential in predicting personal loan eligibility, whereas Logistic Regression gives weight to each feature but does not model complex interactions between them.

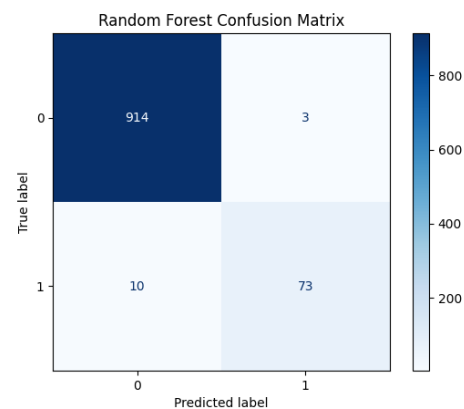
## 3. Confusion Matrix Observations:

- Random Forest's higher recall indicates fewer false negatives (misclassifying eligible applicants as ineligible).
- Logistic Regression had many false negatives, as seen in the recall score, which might have led to missed opportunities for loan approvals.

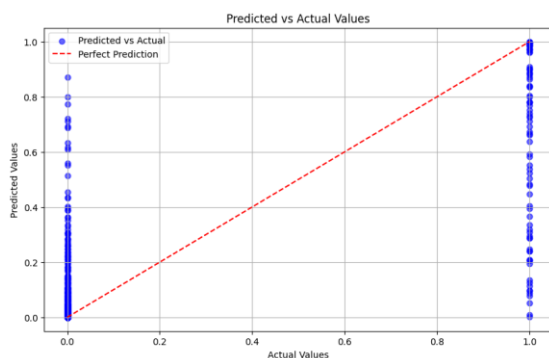
Logistic Regression:



Random Forrest:



Predicted vs Actual:



### 3.Loan Type: Vehicle Loan

#### Sample Dataset:

Age	Gender	Income	Credit Score	Credit History Length	Number of Existing Loans	Loan Amount	Loan Tenure	Existing Customer	State
31	Male	36000	604	487	5	109373	221	No	Karnataka
25	Male	50000	447	386	2	150000	89	No	Karnataka
62	Other	178000	850	503	10	69099	110	Yes	Uttar Pradesh

#### Features and Label

- **Features:** The dataset contains the following features used to predict vehicle loan eligibility:
  - Age, Gender, Income, Credit Score, Credit History Length, Number of Existing Loans
  - Loan Amount, Loan Tenure, Existing Customer, State, City, LTV Ratio (Loan-to-Value Ratio)
  - Employment Profile
- **Label:** Profile Score (converted to binary "Yes" if  $\geq 50$ , "No" if  $< 50$ ).

#### Models Used

1. **Stochastic Logistic Regression:** Selected for its straightforward approach to estimating relationships between eligibility and input features and stochastic due to its large dataset.
2. **Random Forest Classifier:** Chosen for its ability to model complex relationships and generalize well, using ensembles of decision trees.

#### Performance Metrics

- **Logistic Regression:**
  - Accuracy: 0.828
  - Recall: 0.992
- **Random Forest Classifier:**
  - Test Accuracy: 0.986
  - Recall: 0.993

**Observation:** Random Forest again outperformed Logistic Regression in accuracy and recall, making it a better choice for identifying eligible vehicle loan applicants with this dataset.

#### Observations

1. **Model Performance:**

- The Random Forest model achieved higher accuracy (0.986) and recall (0.993) than Logistic Regression, indicating that it is more effective in capturing eligible vehicle loan applicants.
- Logistic Regression had a slightly lower recall (0.992), resulting in a few missed eligible applicants, suggesting Random Forest's decision-making abilities better fit this dataset.

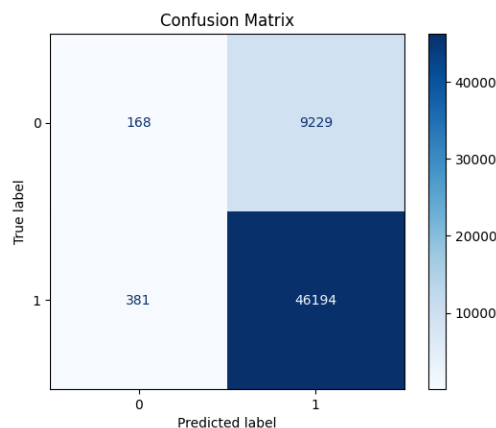
## 2. Feature Importance:

- Random Forest can reveal which features were most influential in loan eligibility, such as Credit Score, Income, and Credit History Length.
- Logistic Regression gives weights to features but may not capture complex interactions as effectively as Random Forest.

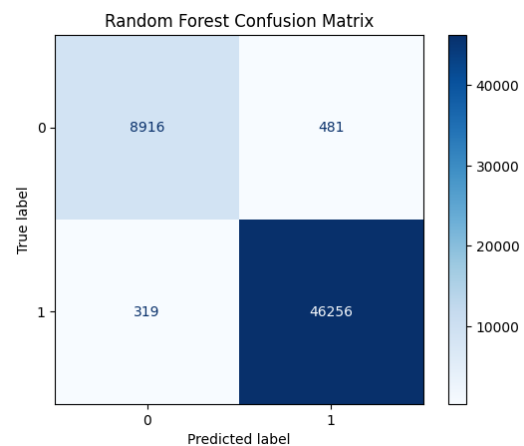
## 3. Confusion Matrix Observations:

- Random Forest's high recall indicates a minimal rate of false negatives (incorrectly classifying eligible applicants as ineligible).
- Logistic Regression showed a similar pattern but with a slightly higher rate of false negatives.

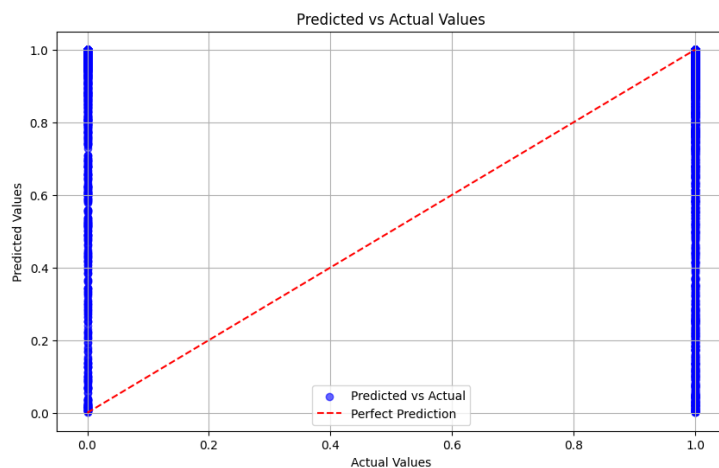
Logistic Regression:



Random forest:



Prediction vs Actual:



#### 4. Other Loans Eligibility Prediction

Data set:

Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term
Male	No	0	Graduate	No	5849	0		360
Male	Yes	1	Graduate	No	4583	1508	128	360
Male	Yes	0	Graduate	Yes	3000	0	66	360
Male	Yes	0	Not Graduate	No	2583	2358	120	360

#### Data Overview

- **Dataset Fields:**
  - **Applicant Demographics:** Gender, Married, Dependents, Education, Self\_Employed, Property\_Area
  - **Financial Information:** ApplicantIncome, CoapplicantIncome, LoanAmount, Loan\_Amount\_Term, Credit\_History
- **Target Label:** Loan\_Status (Predicting loan approval status for other types of loans)

#### Models Used

1. **Logistic Regression**
    - **Accuracy:** 0.8211
    - **Recall:** 0.9881
  2. **Random Forest**
    - **Accuracy:** 0.8130
    - **Recall:** 0.9524
    - **Best Parameters:** n\_estimators=100, max\_depth=10
- **Model Comparison:** **Logistic Regression** outperforms **Random Forest** in both accuracy and recall, making it the more suitable model for predicting eligibility for other loans.

#### Observations

1. **Performance Comparison:**
  - **Logistic Regression** achieved a higher recall (0.9881) and slightly better accuracy (0.8211) compared to Random Forest. This indicates that Logistic Regression is better at correctly identifying eligible loan applicants in the "Other Loans" category.
  - **Random Forest** performed closely but was outperformed by Logistic Regression in both metrics, despite being optimized with the best parameters.
2. **Model Selection:**



- **Logistic Regression** outperformed Random Forest overall, suggesting that it is a more suitable model for this dataset, potentially due to the linear relationships in the features, which Logistic Regression can effectively capture.

### 3. Feature Importance:

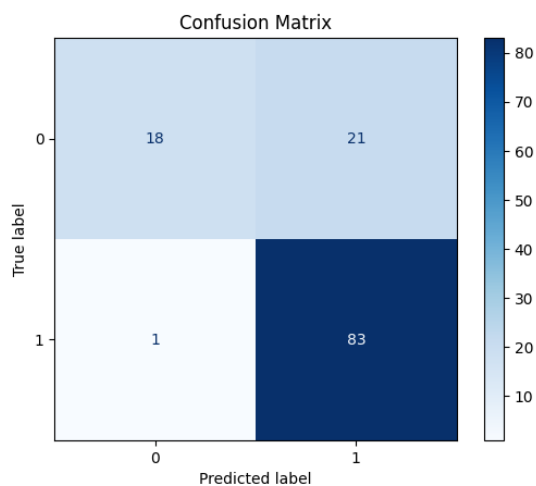
- In Random Forest, analyzing feature importance could reveal which factors have the most influence on the loan eligibility prediction, providing actionable insights for the lending process.

### 4. Error Analysis:

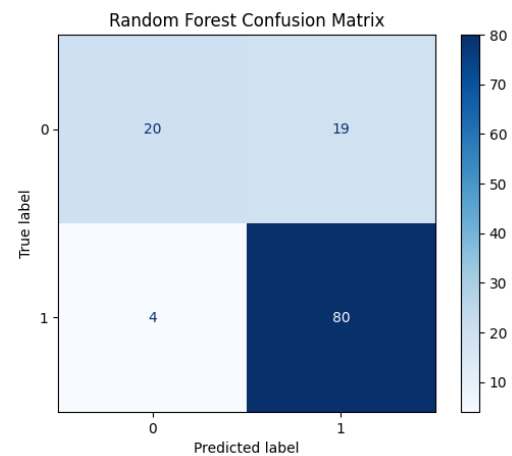
- A **confusion matrix** would provide insights into types of errors (false positives and false negatives) each model is making, helping refine model selection or inform future model adjustments.

Confusion Matrix:

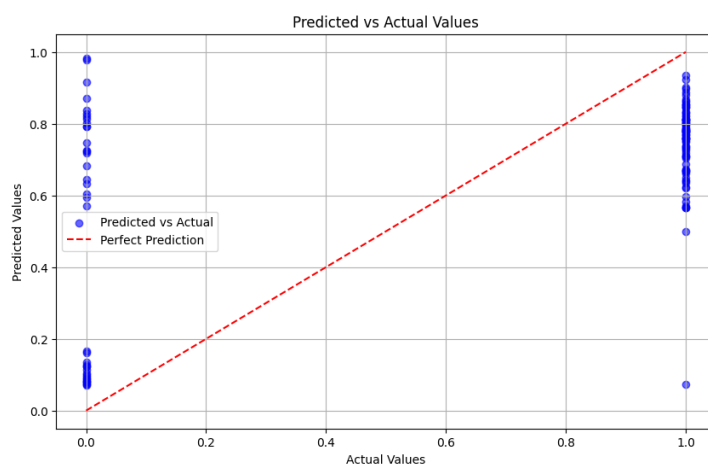
Logistic Regression:



Random Forest:



prediction vs Actual:



## Final Conclusions and Summary

This loan prediction project effectively analyzed eligibility across four loan types—Home Loan, Personal Loan, Vehicle Loan, and Other Loans—by applying Logistic Regression and Random Forest models. Each model was trained with unique dataset features tailored to capture specific characteristics and needs associated with each loan type.

### 1. Model Observations and Performance:

- **Logistic Regression** demonstrated high accuracy and recall in several cases, proving efficient for quick, interpretable results in scenarios with more straightforward feature relationships, like in Other Loans.
- **Random Forest**, with its ensemble approach, generally outperformed Logistic Regression, especially in Personal Loan and Vehicle Loan categories, showing high accuracy and better recall due to its ability to handle complex feature interactions.

### 2. Feature Contributions:

- Each dataset featured a unique set of variables such

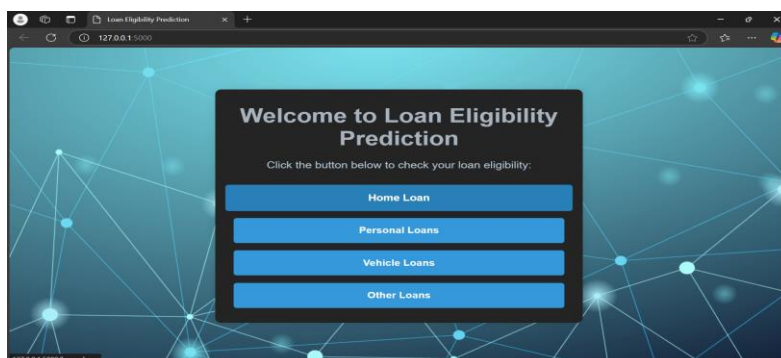
as credit score, applicant income, and credit history length, allowing for a refined analysis that emphasizes the key factors influencing loan eligibility for different loan types.

- The Profile Score threshold of 50 for the Vehicle Loan dataset served as an additional layer for risk evaluation, adding a nuanced control in assessing applicant eligibility.

### 3. Implementation in a Web Application:

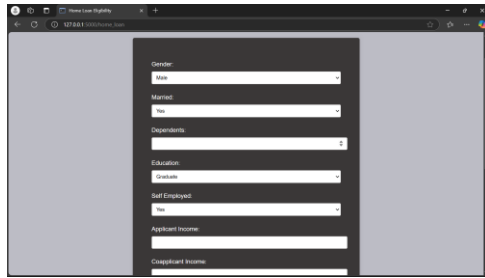
- The project has culminated in a web-based prediction tool, where users can input their data and receive predictions based on pre-trained .pkl models. This application streamlines loan evaluation, enabling quicker assessments with minimal effort.
- Using the site, clients can interact with prediction models directly, receiving tailored loan advice, which enhances transparency and accessibility in the lending process.

Home:



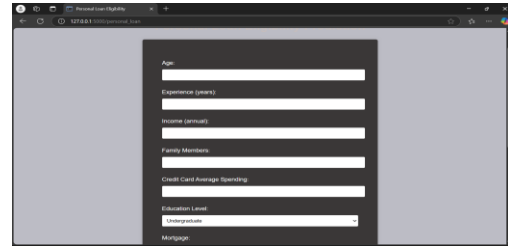
## Interfaces:

### Home Loan:



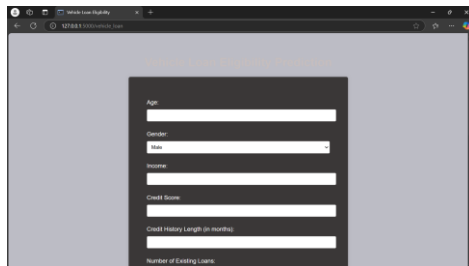
A screenshot of a web browser displaying a form for Home Loan eligibility. The form is titled "Home Loan Eligibility" and includes the following fields: Gender (dropdown menu), Married (checkbox), Dependents (text input), Education (dropdown menu), Self Employed (checkbox), Applicant Income (text input), and Coapplicant Income (text input).

### Personal Loan:



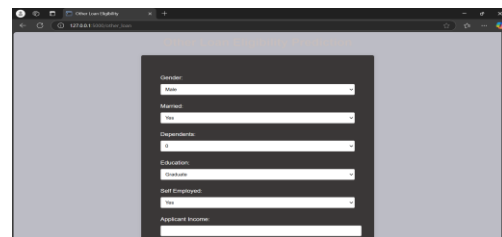
A screenshot of a web browser displaying a form for Personal Loan eligibility. The form is titled "Personal Loan Eligibility" and includes the following fields: Age (text input), Experience (years) (text input), Income (annual) (text input), Family Members (text input), Credit Card Average Spending (text input), Education Level (dropdown menu), and Mortgage (checkbox).

### Vehicle Loan:



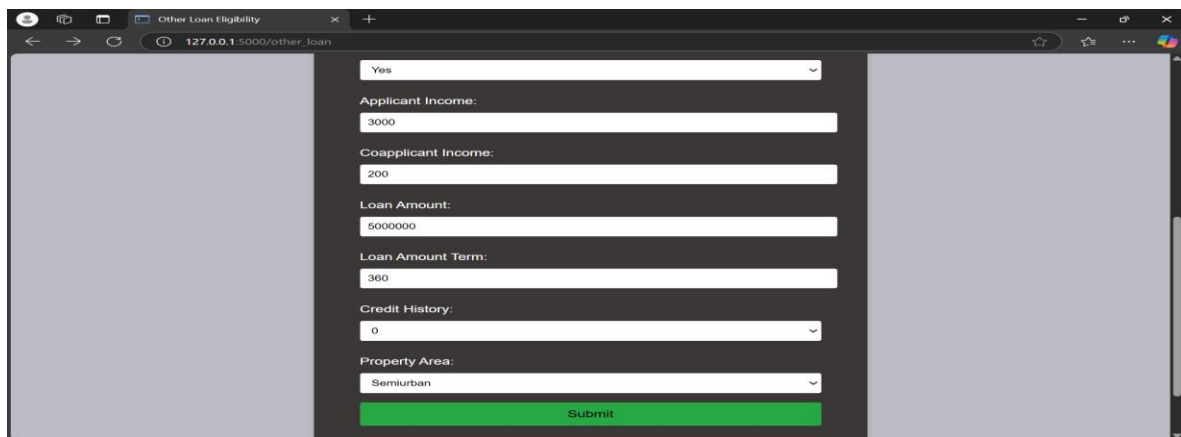
A screenshot of a web browser displaying a form for Vehicle Loan eligibility. The form is titled "Vehicle Loan Eligibility" and includes the following fields: Age (text input), Gender (dropdown menu), Income (text input), Credit Score (text input), Credit History Length in months (text input), and Number of Existing Loans (text input).

### Other Loans:

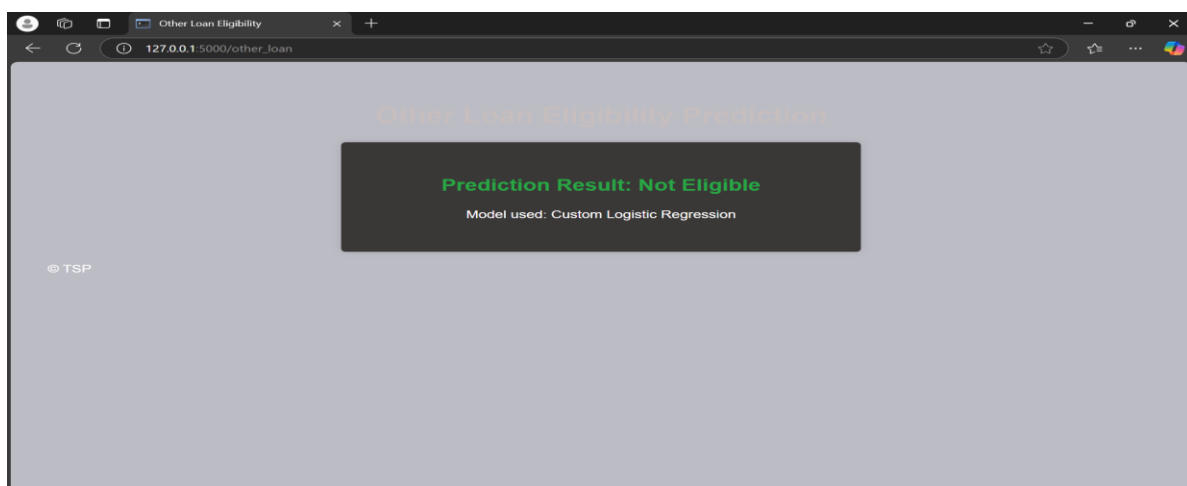


A screenshot of a web browser displaying a form for Other Loans eligibility. The form is titled "Other Loans Eligibility" and includes the following fields: Gender (dropdown menu), Married (checkbox), Dependents (text input), Education (dropdown menu), Self Employed (checkbox), Applicant Income (text input), and Coapplicant Income (text input).

## Overall Implementation and Output:



A screenshot of a web browser displaying the "Other Loan Eligibility Prediction" form. The form is titled "Other Loan Eligibility Prediction" and includes the following fields: Yes (checkbox), Applicant Income: 3000, Coapplicant Income: 200, Loan Amount: 5000000, Loan Amount Term: 360, Credit History: 0, Property Area: Semiurban, and a green Submit button.



A screenshot of a web browser displaying the "Other Loan Eligibility Prediction" output screen. The screen shows the title "Other Loan Eligibility Prediction" and a green box with the text "Prediction Result: Not Eligible" and "Model used: Custom Logistic Regression". The copyright notice "© TSP" is visible in the bottom left corner.