Loan Eligibility Prediction System

Prediction Of Modernized Loan Approval System Based On Machine Learning Approach

Project Introduction

The **Loan Eligibility System** is a data-driven application designed to automate the process of evaluating an individual's eligibility for a loan. The primary objective of this project is to assist financial institutions in making accurate and efficient lending decisions by analyzing key applicant data such as income, employment status, credit history, loan amount, and other financial indicators.

This system leverages data analysis techniques and machine learning models to assess the risk level associated with each applicant and predict loan approval outcomes. By integrating historical data and rule-based logic, the system not only improves decision-making but also minimizes the chances of human error and loan default.

Key features include:

- Data collection and preprocessing from loan applicant datasets.
- Exploratory Data Analysis (EDA) to identify important features influencing loan approval.
- Model building using classification algorithms (e.g., Logistic Regression, Decision Trees, Random Forest).
- Prediction of loan eligibility with accuracy metrics.
- User-friendly interface for data input and result display.

Problem Statement

The company wants to automate the loan eligibility process (real-time) based on customer detail provided while filling out the online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have given a problem identifying the customer segments eligible for loan amounts to target these customers specifically

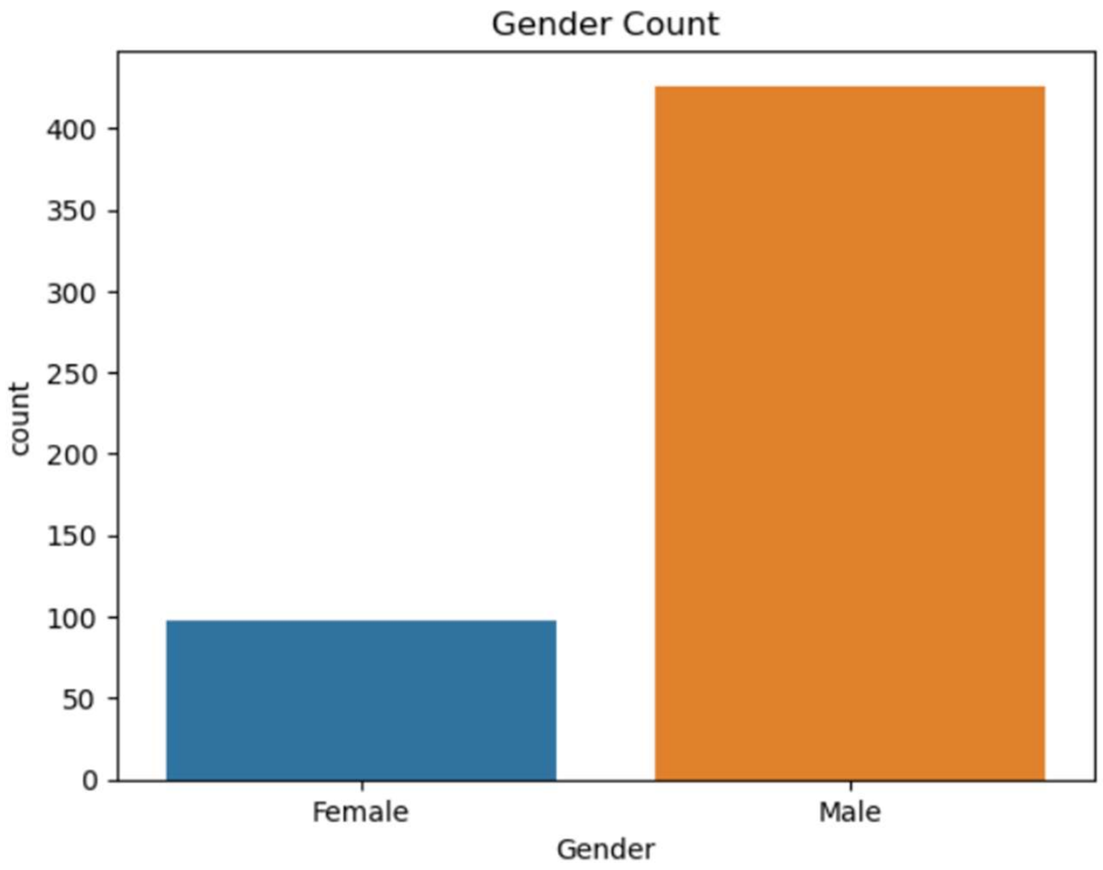
Data Collection

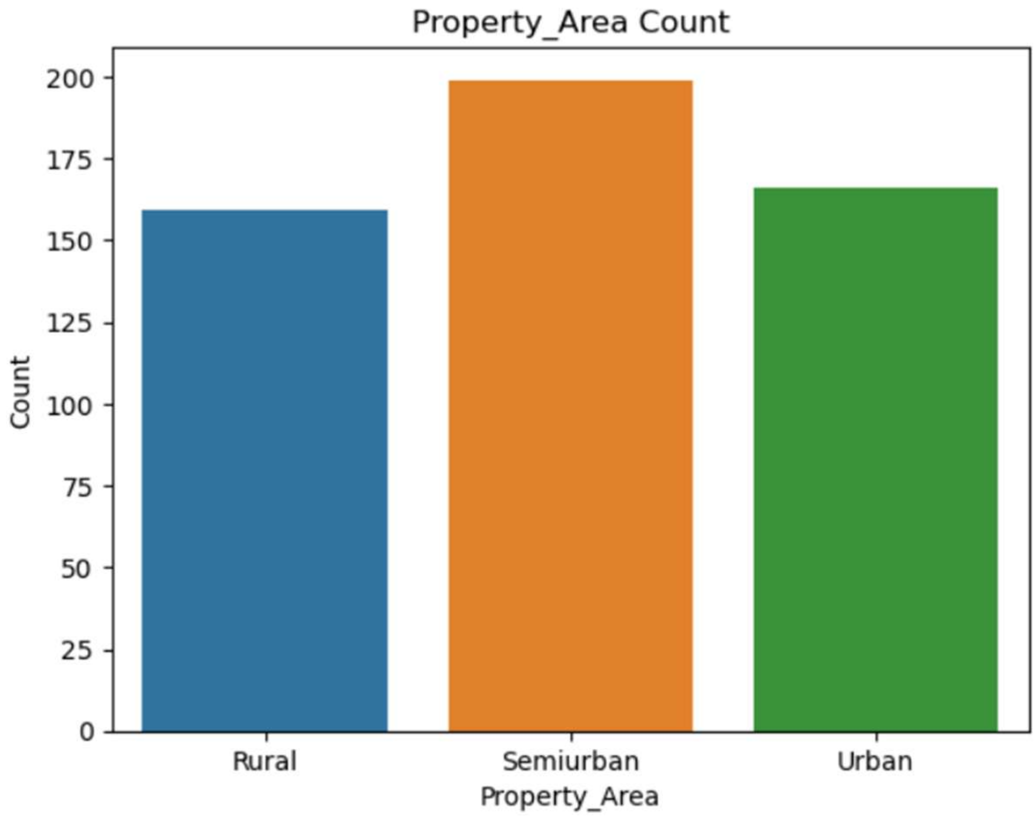
Key Features Collected:

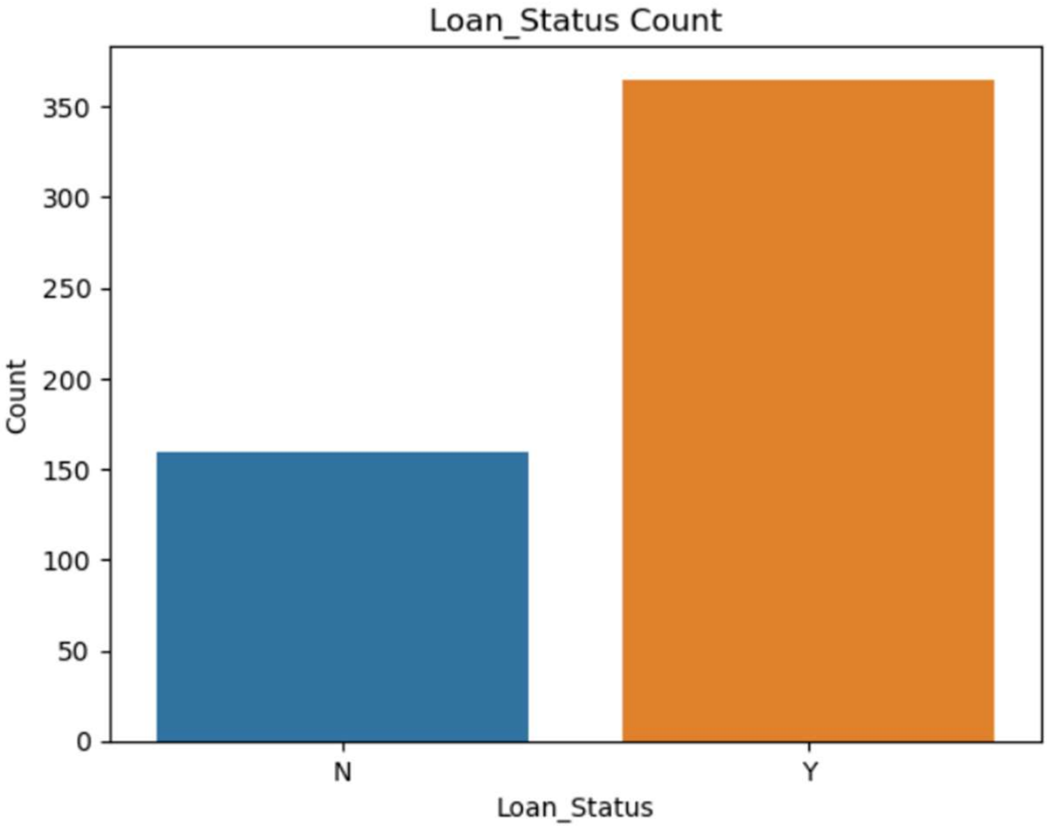
- **ApplicantIncome** – Monthly income of the applicant.
- **CoapplicantIncome** – Income of the co-applicant, if any.
- **LoanAmount** – Amount of loan requested.
- **Loan_Amount_Term** – Term of the loan in months.
- **Credit_History** – Applicant’s credit score (binary: 1 for good, 0 for bad).
- **Gender** – Applicant’s gender.
- **Married** – Marital status.
- **Education** – Educational qualification.
- **Self_Employed** – Employment type.
- **Property_Area** – Area of the property (Urban/Rural/Semiurban).
- **Loan_Status** – Target variable indicating loan approval (Y/N).

Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to understand the underlying patterns, detect anomalies, and identify relationships between variables in the dataset. This step is crucial for preparing the data for model building and for uncovering insights that can influence the loan approval decision.


Gender Count
count
0
50
100
150
200
250
300
350
400
Female
Male
Gender

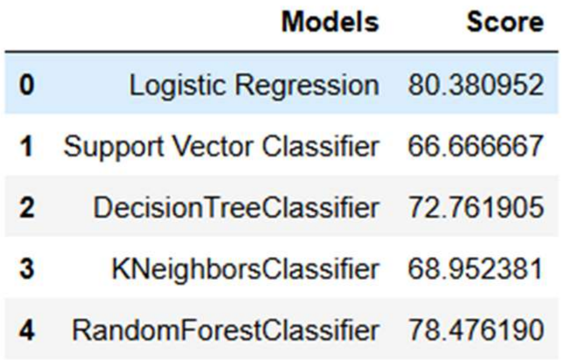

Property_Area Count
Count
0
25
50
75
100
125
150
175
200
Rural
Semiurban
Urban
Property_Area


Loan_Status Count
Count
0
50
100
150
200
250
300
350
N
Y
Loan_Status

Model Selection

The goal of model selection in this project was to identify the best-performing classification algorithm to predict **loan eligibility** (Loan_Status: Yes/No) based on applicant details.

Since the problem is a **binary classification task**, various supervised learning algorithms were tested and compared.



	Models	Score
0	Logistic Regression	80.380952
1	Support Vector Classifier	66.666667
2	DecisionTreeClassifier	72.761905
3	KNeighborsClassifier	68.952381
4	RandomForestClassifier	78.476190

Model Training

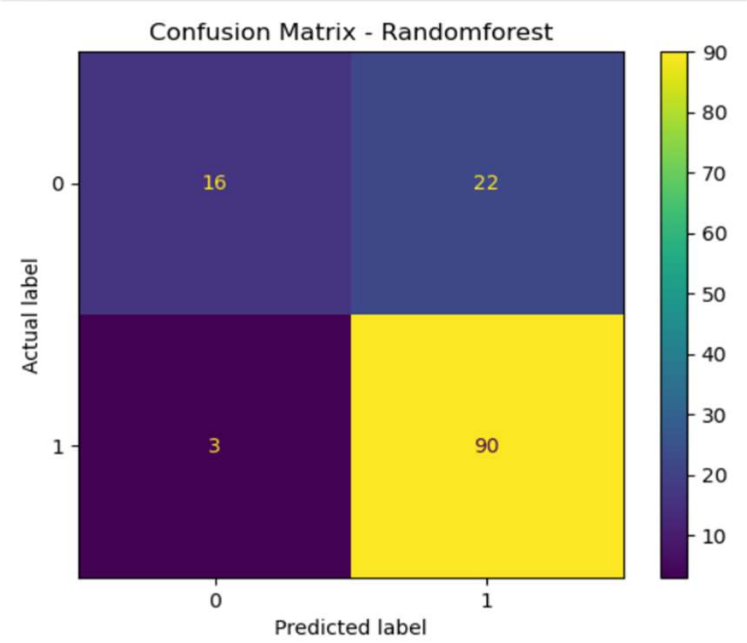
To develop a reliable model for **predicting loan eligibility**, multiple machine learning algorithms were trained using the preprocessed dataset. The goal was to find the model that delivers the highest accuracy in predicting whether a loan should be approved or not.

Algorithm Used: Random Forest Classifier

1. Split dataset into training (80%) and testing (20%)
2. Train Random Forest using training data

Model Evaluation

- After training multiple models, the next step was to evaluate their performance on unseen test data to ensure the model generalizes well and accurately predicts loan eligibility.



Challenges faced

- **Missing values** in key columns (e.g., Credit History)
- **Imbalanced dataset** (more approvals than rejections)
- Categorical data needed proper encoding
- Small dataset size led to **overfitting risk**
- Selecting important features was difficult
- Some models **overfit** on training data

Future Scope

- Use a **larger and more diverse dataset** for better generalization
- Implement **real-time prediction system** using APIs
- Integrate **deep learning models** for improved accuracy
- Add more features like **credit score**, **employment history**, etc.
- Use **automated feature engineering** tools (e.g., Featuretools)
- Continuously **retrain model** with new loan data

Conclusion

- Developed a machine learning model to predict loan approval using applicant information
- **Random Forest** model performed best
- Achieved **84% training accuracy** and **81% testing accuracy**, indicating good generalization
- Preprocessing, feature engineering, and model tuning were key to success
- The system can support lenders in making accurate, data-driven decisions
- Future improvements can include real-time deployment and expanded datasets