


# SuperLender: AI-Powered Loan Default Prediction for Financial Inclusion in Nigeria

---

## Project Overview

This project was developed as part of the **Zindi Nigeria Loan Default Prediction Challenge**, where the goal was to help **digital lending companies** in Nigeria (like *SuperLender*) identify customers most likely to default on their loans.

In emerging economies like **Nigeria**, many individuals lack formal credit histories — making it difficult for financial institutions to make informed lending decisions. This project builds a **Machine Learning (ML)** solution that predicts whether a customer will **repay a loan (Good)** or **default (Bad)**, empowering lenders to make **data-driven, inclusive financial decisions**.

 *This project was officially submitted on Zindi and awarded a Certificate of Completion for successful participation and model submission.*

---

## Problem Statement

SuperLender is a local digital lending platform that wants to improve its **credit risk assessment system**.

The company's goal is to determine, at the time of application, whether a new or returning customer will repay their loan.

The model predicts the **binary target variable** `good_bad_flag`, where:

- 1 → Good (loan repaid)
- 0 → Bad (loan defaulted)

This prediction helps lenders **minimize financial losses**, **reduce default risk**, and **expand access to fair credit** in Nigeria.

---

## Dataset Description

The dataset provided by **Zindi** consists of six CSV files (three for training and three for testing). Each dataset focuses on different aspects of the customer and their loan history.

### Datasets

1. **Demographics (traindemographics.csv)**
  - Customer personal and banking information
  - Includes fields such as `birthdate`, `bank_account_type`, `employment_status_clients`, etc.
2. **Performance (trainperf.csv)**
  - Details about the specific loan performance being predicted
  - Key variable: `good_bad_flag` (target)

### 3. Previous Loans (trainprevloans.csv)

- History of all previous loans taken by each customer

👉 All datasets were merged using the unique key **customerid** to create a comprehensive view of each customer.

---

### ⚙️ Data Preprocessing

To ensure data quality and modeling efficiency, the following steps were performed:


- Merged datasets on customerid
- Removed duplicate and irrelevant features
- Handled missing values (NaN imputation and dropping sparse columns)
- Created new features:
  - age (calculated from birthdate)
  - loan\_duration (closeddate - approveddate)
  - repayment\_ratio (totaldue / loanamount)
- Encoded categorical columns using Label Encoding
- Scaled numerical features using StandardScaler

---

### 🧠 Model Development

A range of models were tested to find the most effective one for this financial risk prediction task:

Model	Description	Result
Logistic Regression	Baseline classifier	Good baseline
Random Forest	Robust tree-based model	Improved accuracy
XGBoost	Gradient boosting model	High predictive power
CatBoost	Handles categorical data efficiently	Great recall

**Stacking Ensemble** Combines the above models  **Final Model**

The **Stacking Ensemble Model** (Random Forest + XGBoost + CatBoost with Logistic Regression as meta-learner) achieved the best balance between **precision** and **recall**, minimizing both false positives and negatives.

---

### 📊 Model Evaluation

Evaluation Metrics used:

- Accuracy
- Precision
- Recall
- F1 Score
- ROC-AUC

☑ The model achieved strong performance on the validation set, effectively identifying risky borrowers while maintaining fairness in prediction outcomes.

## 🚩 Final Model and Predictions

- Trained final model on full dataset using optimal hyperparameters.
- Generated predictions for the Zindi test set following the required submission format:

customerID,Good\_Bad\_flag

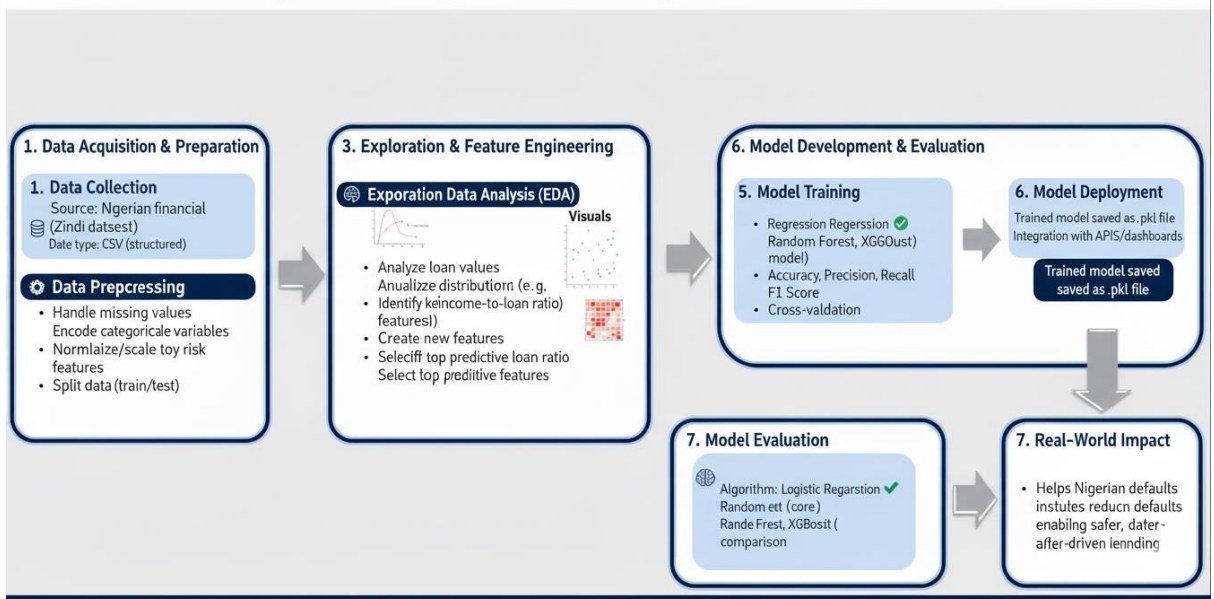
12345667,1

43423156,0

54325779,0

- Saved artifacts:
  - best\_model.pkl (Trained Model)
  - scaler.pkl (Feature Scaler)
  - final\_submission.csv (Predictions)

## SuperLender – ML Pipeline Overview



---

## 📁 Repository Structure

```
SuperLender-AI-Loan-Default-Prediction/
|
├─ data/
|   ├── traindemographics.csv
|   ├── trainperf.csv
|   ├── trainprevloans.csv
|   ├── testdemographics.csv
|   ├── testperf.csv
|   └── testprevloans.csv
|
├─ notebooks/
|   ├── data_exploration.ipynb
|   └── model_training.ipynb
|
├─ src/
|   ├── data_preprocessing.py
|   ├── train_model.py
|   └── evaluate_model.py
|
├─ models/
|   ├── best_model.pkl
|   └── scaler.pkl
|
├─ requirements.txt
├─ README.md
└─ LICENSE
```

---

## 🔧 Tech Stack

### Languages & Libraries:

- Python
- Pandas, NumPy
- Scikit-learn
- XGBoost, CatBoost

- Matplotlib, Seaborn
- 

## Results

- **Model Type:** Stacking Ensemble
  - **Performance:** Balanced precision and recall with minimal misclassification
  - **Top Features:**
    - Employment status
    - Loan amount
    - Loan duration
    - Total due
    - Repayment ratio
- 

## How to Run the Project


1. **Clone the repository**
  2. `git clone https://github.com/Rupeshbhardwaj002/SuperLender-Smarter-Loans-with-Data-Driven-Decisions/tree/main.git`
  3. `cd SuperLender-Loan-Default-Prediction`
  4. **Install dependencies**
  5. `pip install -r requirements.txt`
  6. **Run preprocessing and training**
  7. `python src/data_preprocessing.py`
  8. `python src/train_model.py`
  9. **Generate predictions**
  10. `python src/evaluate_model.py`
  11. **Output file:**
    - `final_submission.csv`
    - `best_model.pkl`
- 

## Learnings & Real-World Impact

Through this project, I learned:

- How to work with **real multi-table financial data**

- Handling **missing values and categorical encoding**
- Designing **stacking ensemble architectures** for performance gains
- The importance of **responsible AI in lending** — improving **financial inclusion** for Nigerian citizens through fair, data-backed credit scoring.

 This project demonstrates how AI can help bridge the gap between underserved individuals and accessible financial services in Africa.

---

## Certificate of Completion

✓ Successfully submitted on **Zindi Africa** and awarded a **Certificate of Completion** for active participation in the *Loan Default Prediction Challenge*.

# Zindi Certificate

*rupesh002 has participated in the following competitions:*

## Loan Default Prediction Challenge

<https://zindi.africa/competitions/data-science-nigeria-challenge-1-loan-default-prediction>

*Can you predict who will default on a loan?*  
11 October 2018–1 January 1970

currently ranked


**367 out of 459**

To verify Please visit - <https://zindi.africa/users/rupesh002/competitions/certificate>

---

## License & Credits

- **License:** MIT
  - **Dataset & Problem Source:** Zindi Africa
  - **Author:** Leon Hartmann
- 

 "Empowering Financial Inclusion through Data-Driven Intelligence."

---