Model Research

1. Objective of the Model

The goal of this machine learning model is to predict loan eligibility for users based on financial and personal information extracted from identity documents (like Aadhar, PAN) and bank statements from the last 6 months. The model helps automate creditworthiness checks, improve consistency, and provide transparency using explainability tools.

2. Data Sources and Feature Engineering

Data Sources:
- Aadhar Card – Name, DOB, Aadhar Number
- PAN Card – Name, PAN Number, Father's Name
- Bank Statements (6 months) – Extracted from uploaded PDFs/images

Key Features Extracted from Bank Statements:

- avg_monthly_balance: Average closing balance
- total_credits_last_6_months: Sum of all credited amounts
- total_debits_last_6_months: Sum of all debited amounts
- max_single_credit / debit
- salary_detected (boolean)
- account_stability_score
- avg_monthly_credits / debits

Other Standard Features Used:

- Age, Gender, Marital Status, Dependents
- Education, Employment Type
- Credit Score, Loan Amount, Loan Term
- Property Area

Derived Features:

- debt_to_income_ratio = estimated_emi / avg_monthly_credits
- loan_amount_per_income = loan_amount / avg_monthly_credits
- credit_score_normalized = (credit_score - 300) / 550
- bank_health_score (custom metric)

3. Model Selection

Algorithms Evaluated:
- Logistic Regression
- Random Forest
- XGBoost
- SVC (optional)

4. Training Pipeline

- Dataset Size: 5000 synthetic + test samples
- Normalized transactions from bank data
- SMOTE used for class imbalance
- Cross-validation applied

5. Performance Metrics

- Accuracy: ~89%
- Precision: High (avoid false approvals)
- Recall: Moderate (capture real approvals)
- AUC-ROC: > 0.92

6. Model Explainability

Used SHAP (SHapley Additive Explanations):
- Feature-wise attributions
- Top 5 positive/negative contributing features
- Human-readable justifications

7. Model Deployment & Serving

- Saved model files: best_model.pkl, scaler.pkl, encoders.pkl, explainer.pkl
- FastAPI backend loads model on startup
- /predict endpoint receives structured data or extracted data
- Real-time predictions and explanations returned

Summary

Final model reasons over bank behavior and personal data, predicts eligibility with >89% accuracy, provides clear explanations, and is production-ready via REST APIs.