LOAN DEFAULT PREDICTION SYSTEM - COMPREHENSIVE PROJECT DESCRIPTION

(GROUP 5)

1. PROJECT OVERVIEW

This Streamlit-based web application is designed to help financial institutions predict the likelihood of loan defaults using machine learning. The system follows a structured end-to-end pipeline, from data ingestion to model deployment, providing actionable insights for risk assessment.

2. KEY FEATURES

A. Data Management

- CSV Upload & Exploration: Users can upload loan data or use the default dataset.
- Missing Value Handling: Imputation for numerical (median) and categorical (mode) features.
- Automated Preprocessing:
  - ❖ Standardization (for numerical features) → Ensures equal feature weighting.
  - ❖ One-Hot Encoding** (for categorical features) → Converts text categories into model-friendly numerical values.

B. Feature Selection & Engineering

- Correlation Analysis**: Identifies features strongly linked to loan defaults.
- Best Subset Selection** (Sequential Feature Selection):
  - ❖ Uses Logistic Regression as a base model.
  - ❖ Evaluates feature combinations via **cross-validation** (5-fold CV).
  - ❖ Optimizes for accuracy, precision, recall, or F1-score.

C. MODEL TRAINING & EVALUATION

- Algorithm: Random Forest Classifier (ensemble method for robustness).
- Hyperparameter Tuning:
    - ❖ Adjustable number of trees, max depth, and min samples split.
- Performance Metrics:
    - ❖ Accuracy, Precision, Recall, F1-Score.
    - ❖ Confusion Matrix (visualizes true/false positives/negatives).
    - ❖ Feature Importance (identifies key risk factors).

D. INTERACTIVE PREDICTION

- Real-Time Risk Assessment:
    - ❖ Users input loan applicant details (income, credit score, loan amount, etc.).
    - ❖ The model returns default probability (Low/Medium/High Risk).
- Risk Factor Breakdown:
    - ❖ Explains why an applicant is flagged as risky (e.g., "Low credit score (<580)").
- Diagnostic Tools:
    - ❖ Shows raw probabilities for debugging.
    - ❖ Adjustable risk thresholds (default: >30% = High Risk).

E. BUSINESS INSIGHTS & REPORTING

- Model Interpretation:
    - ❖ Summarizes key findings (e.g., "High debt-to-income ratios increase default risk").
    - ❖ Provides recommendations for loan officers (approval/interest rate adjustments).
- Limitations & Future Improvements**:
    - ❖ Discusses model constraints (e.g., "Does not account for macroeconomic factors").
    - ❖ Suggests enhancements (e.g., "Try XGBoost for better performance").

3. TECHNICAL IMPLEMENTATION

A. Machine Learning Pipeline

1. Data Loading → `load_data()`

- Reads CSV, drops irrelevant columns (`ID`, `dtir1`), caches for efficiency.

2. Preprocessing → `create_ preprocessor()`

- Numerical Features: Median imputation + StandardScaler.
- Categorical Features: Mode imputation + OneHotEncoder.
- ColumnTransformer: applies transformations in parallel.

3. Feature Selection → `Feature_Selection_page()`

- Uses forward selection to pick optimal features.

4. Model Training → `Model_Selection_And_Training_page()`

- Random Forest with customizable hyperparameters.
- Cross-validation (5-fold) prevents overfitting.

5. Prediction → `Interactive_Prediction_page()`

- Preprocesses input → generates risk score → explains decision.

B. Key Libraries Used

LIBRARY	**PURPOSE**
Streamlit	Web app framework
Pandas	Data manipulation
Scikit-Learn	ML models & preprocessing
Matplotlib/Seaborn	Visualizations
Pickle	Model serialization

C. Performance Optimization

- Caching (`@st.cache_data`): Speeds up repeated computations.
- Modular Design: Separates data, model, and UI logic for maintainability.
- Persistent Artifacts: Saves models/preprocessors to disk (`DATA_DIR`).

4. BUSINESS IMPACT

A. For Loan Officers

- Instant Risk Scoring: Approve/reject loans faster.
- Risk-Based Pricing: Adjust interest rates based on predicted default probability.
- Flagging High-Risk Cases: Manual review for borderline applicants (e.g., 40-60% risk).

B. For Risk Managers

- Portfolio Analysis: Identify high-risk loan segments.
- Model Monitoring: Track performance over time.
- Regulatory Compliance: Transparent, data-driven decisions.

5. LIMITATIONS & FUTURE WORK

A. Current Limitations

1. Data Dependency: Requires high-quality historical loan data.

2. Black-Box Nature: Random Forests are less interpretable than logistic regression.

3. Economic Factors: Doesn’t account for recessions or policy changes.

B. Planned Improvements

1. Alternative Models: Test XGBoost or Neural Networks.

2. More Features: Add employment history or macroeconomic indicators.

3. Dynamic Thresholds: Auto-adjust risk thresholds based on market conditions.

6. TEAM & DEPLOYMENT

SN	TEAM MEMBER	STUDENT ID	ROLE	CONTRIBUTION
1	Kingsley Sarfo	22252461	Project Lead	App Design & Preprocessing
2	Francisca Manu Sarpong	22255796	Deployment	Streamlit Cloud Integration
3	George Owell	22256146	Model Evaluation	Cross-Validation & Metrics
4	Barima Owiredu Addo	22254055	UI/Testing	Prediction Interface
5	Akrobettoe Marcus	11410687	Feature Engineering	Best Subset Selection

7. CONCLUSION

This system provides a scalable, automated solution for loan default prediction, combining machine learning best practices with an intuitive interface. By identifying high-risk applicants early, financial institutions can reduce losses while maintaining fair lending practices.