Ryan_Report — FinTech & Risk Assessment (Loan Default/Approval Risk)

Date: February 20, 2026

1. Introduction (Business Problem)

InclusiFinance is a FinTech startup offering micro-loans to borrowers who may be underserved by traditional banks. The business must balance two competing risks: (a) being too strict and denying credit to deserving applicants (hurting inclusion), and (b) being too lenient and approving high-risk borrowers (causing losses). This project builds and evaluates machine-learning models that estimate borrower risk so the Head of Risk Management can set an approval policy.

Target Definition and Important Limitation

The dataset used is a Loan Approval dataset with label **loan_status** (Approved/Rejected). Because a true default flag is not present, **Rejected** is treated as a proxy for high-risk/default-prone applicants. In real deployment, the preferred target would be an observed default/late-payment outcome after loan issuance.

2. Data & Methodology

Data source: Kaggle “Loan Approval Prediction Dataset” (https://www.kaggle.com/datasets/architsharma01/loan-approval-prediction-dataset). Dataset size: 4,269 rows, 13 raw columns (one identifier + applicant features + label).
Preprocessing: trimmed column/value whitespace; encoded categorical features (education, self_employed); created a financial ratio feature **loan_income_ratio = loan_amount / income_annum**; removed non-predictive identifier (loan_id); standardized numeric features for models sensitive to scale (Logistic Regression and MLP).
Class imbalance handling: used **stratified train/test split** to preserve class proportions and reported metrics robust to imbalance (Precision/Recall/F1/ROC-AUC).

3. Exploratory Data Analysis (EDA) Highlights

Key visual findings: (1) class balance is moderately imbalanced (Approved > Rejected); (2) CIBIL score strongly separates Approved vs Rejected applicants; (3) loan_income_ratio and loan_term provide additional signal; (4) asset variables provide a stability/cushion signal. These observations motivated using both interpretable baselines and non-linear ensemble models.

4. Models Trained

Baseline (interpretable): Logistic Regression.
Tree-based tabular models: Random Forest and Gradient Boosting.
Neural network: Multi-Layer Perceptron (MLP) with two hidden layers.

5. Results (Test Set)

Model	Accuracy	Precision	Recall	F1	ROC-AUC
Logistic Regression	0.9157	0.9088	0.8638	0.8857	0.9744

Random Forest	0.9965	1.0000	0.9907	0.9953	0.9995
MLP (Neural Net)	0.9684	0.9654	0.9505	0.9579	0.9939
Gradient Boosting	0.9941	0.9938	0.9907	0.9922	0.9990

Takeaway: Tree ensembles (Random Forest / Gradient Boosting) substantially outperform Logistic Regression and MLP on this dataset. Random Forest achieves the highest raw metrics; Gradient Boosting is extremely close and is often preferred in credit-risk practice due to stability and probability quality. Logistic Regression remains valuable for interpretability and regulatory explanation.

Interpretability & Key Risk Drivers

Interpretability was assessed via Logistic Regression coefficients and ensemble feature importance. Across models, **CIBIL score** is the dominant signal (higher score → lower risk). Other consistent drivers include **loan_term**, **loan_income_ratio**, and asset-related features (bank/luxury/commercial/residential assets). For explaining individual decisions in production, a FinTech team would typically use feature attributions (e.g., SHAP) with boosting models.

Operational Trade-offs (Training Time & Complexity)

Operationally, Logistic Regression trains fastest and is simplest to retrain. Gradient Boosting and Random Forest require more compute but remain practical for this dataset size; MLP requires scaling and hyperparameter tuning and is less explainable. In a regulated setting, the marginal performance gain must be weighed against explainability and audit requirements.

6. Thin-File / No-CIBIL Stress Test

To simulate underserved borrowers with limited credit history, we retrained Gradient Boosting after removing CIBIL score. Performance dropped sharply (Accuracy 0.607, Recall 0.124, ROC-AUC 0.586), showing that this dataset's decisions depend heavily on credit score. Business implication: InclusiFinance should not rely solely on this feature set for credit-invisible borrowers; it should either collect alternative data (cash-flow stability, rent/utility payments, bank transaction signals) or adopt **progressive lending** (small starter loans to build internal behavioral credit).

7. Recommendation to Head of Risk Management

Recommendation: Deploy **Gradient Boosting** as the primary model for risk scoring due to near-best performance and strong operational reliability, and maintain **Logistic Regression** as an interpretable benchmark for audits and policy communication.
Policy: Use model probabilities to create risk bands (Approve / Review or small starter loan / Reject), and tune the decision threshold based on portfolio loss tolerance.
Next steps: Evaluate calibration, add fairness checks, and expand features for credit-invisible borrowers using alternative data sources.