

Loan Approval Prediction for a FinTech Lender

Executive Summary

1. Introduction

This project analyses a historical loan application dataset (4,269 records, 13 features) to support a FinTech lender in making more data-driven and consistent loan approval decisions. Each record contains demographic and financial information (dependents, education, employment type, income, requested loan, credit score, asset values) and an outcome label: *Approved* (62.2%) or *Rejected* (37.8%).

The objectives are to:

- Build a predictive model that estimates the probability of loan approval for new applications.
- Identify which financial factors are most strongly associated with approval.
- Provide business recommendations and highlight key ethical considerations.

2. Methodology

2.1 Data preparation and feature engineering

Key steps:

- **Cleaning and split.** No missing or duplicate records were found. The identifier (`loan_id`) was dropped. Data were split into 80% training and 20% test using stratified sampling.
- **Encoding and scaling.** The target was encoded as `Approved = 1`, `Rejected = 0`. Categorical predictors (education, self-employed) were one-hot encoded. Continuous predictors (income, loan amount, term, credit score, asset values) were standardised.
- **Feature engineering.** To reflect repayment capacity and leverage, the following ratios were added: loan-to-income, loan-to-total-assets, income per dependent, plus an aggregate total-assets feature (sum of residential, commercial, luxury and bank assets).

2.2 Models evaluated

Four supervised classification models were trained and compared:

- **Logistic Regression** – linear baseline, fast and interpretable.
- **Decision Tree** – non-linear, rule-based model.
- **Random Forest** – ensemble of decision trees on bootstrapped samples.
- **Gradient Boosting** – sequential ensemble of shallow trees.

Because approvals are more frequent than rejections, tree-based models used class balancing so that misclassifying a rejected loan receives higher penalty. Performance was evaluated on the held-out test set using accuracy, precision, recall, F1-score and Area Under the ROC Curve (AUC), plus confusion matrices to examine false approvals and false rejections.

3. Key Findings

3.1 Model performance

Table 1 compares the baseline Logistic Regression with the best-performing ensemble model, Random Forest, on the test set (854 applications).

Model	Accuracy	Precision (Approved)	Recall (Approved)	ROC AUC
Logistic Regression	0.924	0.955	0.921	0.974
Random Forest	0.982	≈0.985	≈0.987	0.999

Table 1: Test-set performance of baseline and best model.

Logistic Regression already achieves strong performance (92.4% accuracy, AUC 0.974), but still makes a noticeable number of false approvals and false rejections.

The Random Forest is clearly superior, with 98.2% accuracy and AUC of 0.999, indicating near-perfect discrimination between good and bad applications. Its confusion matrix shows only a small number of false rejections and

false approvals (both around 1–3% of the test set). Gradient Boosting performs almost as well ($AUC \approx 0.998$), while the single Decision Tree trails the ensembles on all key metrics.

3.2 Important drivers of approval

Exploratory analysis and model behaviour highlight several key drivers:

- **Credit score (CIBIL).** Approved customers have substantially higher scores than rejected ones; this is one of the most informative predictors.
- **Income and assets.** Higher annual income and higher asset values strongly correlate with approval, reflecting repayment capacity and available collateral.
- **Leverage ratios.** High loan-to-income and loan-to-total-assets ratios are associated with higher rejection rates, reflecting the risk of very large loans relative to income and assets.
- **Household burden.** Lower income per dependent is linked to higher rejection probability, indicating financial pressure from larger households.

Overall, the ensembles combine credit history, income, assets and engineered ratios to differentiate low-risk from high-risk applicants in a way that is consistent with standard credit risk principles.

4. Recommendations and Business Insights

4.1 Recommended model

Given its predictive performance and robustness, the **Random Forest** model is recommended as the primary engine for automated pre-screening of loan applications. It offers:

- Very high accuracy and AUC, meaning most risky customers are correctly rejected while good customers are approved.
- Low false positive rate, limiting the number of loans granted to high-risk applicants.
- Low false negative rate, reducing lost revenue from unnecessarily rejected customers.

In production, the model's probability scores can be mapped to simple business rules, for example:

- Automatically approving applications above a high probability threshold.
- Automatically declining applications below a low threshold.
- Sending borderline cases to human credit officers for manual review.

4.2 Strategic insights for the FinTech firm

The analysis suggests several actionable insights:

- Maintaining good **credit scores** should be central to customer education, as this is a key determinant of approval.
- Internal risk policies should explicitly consider **loan-to-income** and **loan-to-assets** ratios, as these align with what the model learns from the data.
- Predicted risk scores can support differentiated pricing or loan limits, with human oversight for atypical cases.

5. Ethical Considerations (Bonus)

Deploying automated loan approval models raises important ethical and regulatory questions:

- **Bias and fairness.** Even without explicit sensitive attributes (e.g., gender, ethnicity), financial variables can act as proxies. Before deployment, the model should be audited for disparate impact across demographic groups where legally permitted, with mitigation strategies (fairness-aware training, adjusted thresholds, policy overrides) considered.
- **Transparency and explainability.** Ensemble models like Random Forest are not inherently transparent. To maintain trust and meet regulatory expectations, the firm should provide clear explanations of key decision factors (e.g., low credit score) and use tools such as feature importance or SHAP analysis for case-level explanations.
- **Human oversight and privacy.** The model should support, not fully replace, human judgement, especially for borderline cases or customers with limited history. All personal and financial data must be processed in compliance with data protection regulations, with strict access control and logging.