Credit Risk Management Using Machine Learning

Objective The goal is to build a classification model that predicts the likelihood of a customer defaulting on a loan. Such a system enables fintech platforms to automate credit decisions and offer faster, data-driven services—making them more agile than traditional banks.

Project Overview This project explores how modern fintech companies leverage machine learning to assess the creditworthiness of individuals. Using the publicly available German Credit dataset (with over 1000 customer records), we built a predictive model to classify customers as credible (1) or non-credible (0) based on their financial behavior and demographic attributes.

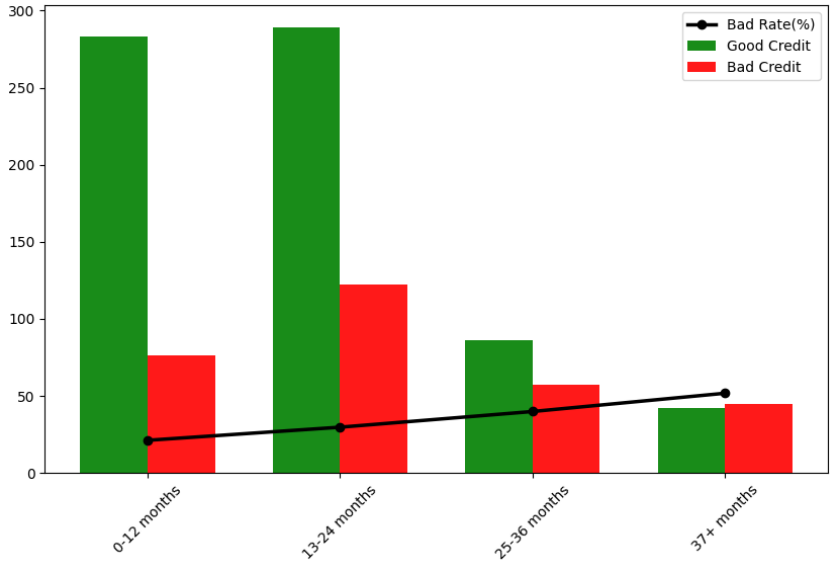
Dataset Summary

- **Total Records**: 1000 customers
- **Class Distribution**:
  - 700 were labeled as credible (no default)
  - 300 were labeled as non-credible (default)
- **Features Used**:
  - Age (binned into logical segments)
  - Duration of Credit
  - Account Balance
  - Number of Credits
  - Sex/Marital Status

```
Feature Importance:
                      Feature  Importance
0                   Age_years    0.405263
1  Duration_of_Credit_monthly    0.264236
2             Account_Balance    0.186974
4          Sex_Marital_Status    0.082489
3  No_of_Credits_at_this_Bank    0.061037
```

Key Insights from EDA

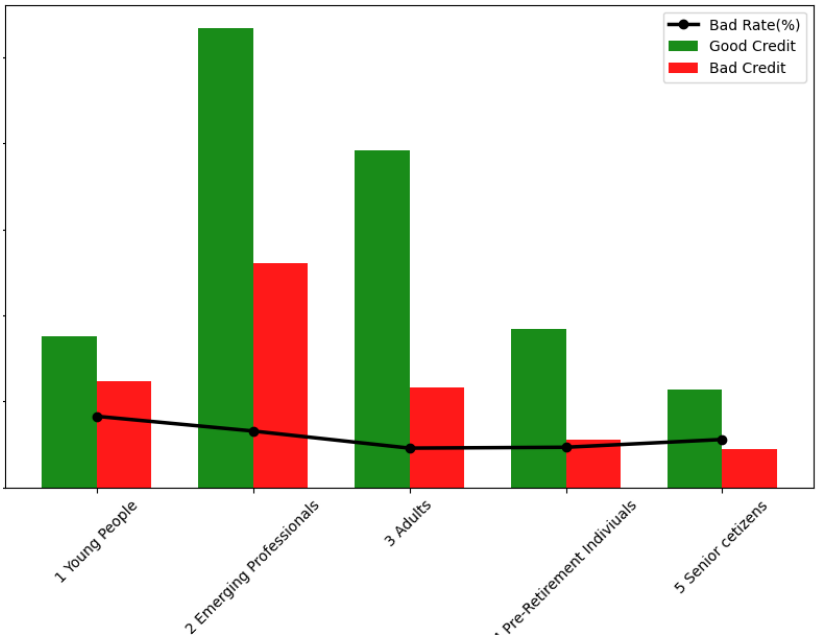
1. **Credit Duration & Risk**:
   - Short-term loans (0–12 months) were the most popular (359 loans, 77%).
   - However, as duration increased, the bad credit rate also rose:
     - 0–12 months: 21.17%
     - 13–24 months: 29.68%
     - 25–36 months: 39.86%
     - 37+ months: 51.72%
   - Insight: Long-term commitments increase borrower risk exposure, suggesting a need for stricter evaluation on longer-duration loans.
     The model would take an assumption of shorter durations to be much safer.



(Figure – Duration-wise Default Rates)

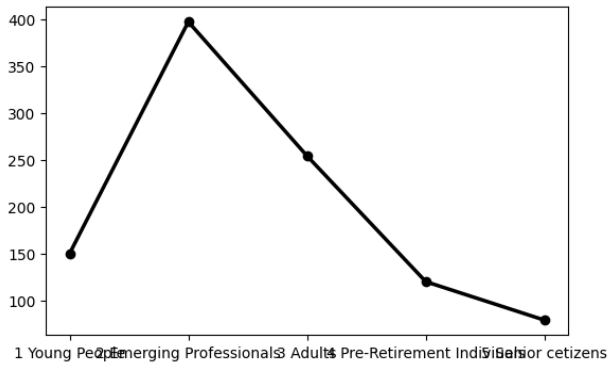
2. **Age Group Risk Analysis**:

   - The customer base was segmented into five age groups:
     - Young People (<25)
     - Emerging Professionals (25–34)
     - Adults (35–44)
     - Pre-Retirement Individuals (45–54)
     - Senior Citizens (55+)
   - Young People had the highest bad credit rate (41.33%), despite a low average loan amount ($2970.73).
   - Adults and Pre-Retirement individuals showed more financial discipline (bad rates ~22–23%).

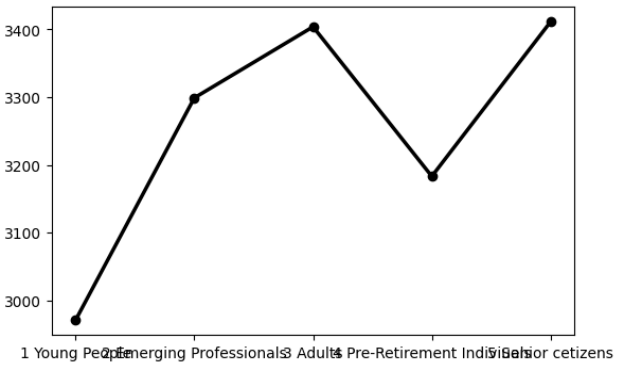


(Fig- Age Class wise Bad Credit Rates)

Young People and Emerging Professionals form a large and fast-growing customer segment, but they also show above-average default rates. Ignoring them would limit growth, but serving them recklessly increases risk.



(Figure – Number of customers per Age Group)



(Figure – Net Credit per Age Group)

Strategy for Real-World Adoption for Young People:

Introduce Safer Credit Products

- Launch starter BNPLs or small-ticket loans with low limits.
- Allow credit limit upgrades based on repayment history.
- Offer secured credit (e.g., backed by savings or deposits) for customers with a high probability of default.

  Example: A user may need to maintain a minimum balance.

Incentivize Good Behaviour

- Use gamified rewards for on-time payments
- Send timely nudges or reminders to reduce forgetfulness
- Offer lower interest rates or cashbacks for consistent repayment

Use Granular Risk Scoring

- Use machine learning to assign probability-based risk scores (not just approve/reject)
- Score applicants within the segment (e.g., one 24-year-old may be less risky than another based on account balance, duration, etc.)

This blended approach preserves growth from a large segment while applying data-driven safeguards to reduce defaults. Over time, it also builds customer loyalty and better financial habits—a win-win for both the lender and the borrower.

3. **Account Balance & Credit Risk**:
    - A strong negative correlation was observed between account balance and bad credit rate:
        - No Checking Account: 49.27%
        - No Balance: 39.03%
        - Some Balance: 22.22%
        - Great Balance: 11.68%
    - Insight: Positive cash flow and balance history are strong indicators of creditworthiness. The model will give a low probability of default to people with a Balance.
    - Additional Insight: Offering credit-builder products to customers with "no checking account" or "no balance" may improve their financial standing and lower risk in the future.

(Figure – Bad Rates vs Presence of Account and Balance)

4. **Demographic & Behavioral Patterns**:
    - Male and single individuals had slightly higher default rates than their married counterparts (based on model features and test observations).
    - Insight: Marital status combined with credit history can help fine-tune risk assessments.

5. **Loan Product Design**:
    - Based on observed risk patterns, fintechs could:
        - Promote short-term credit products with lower limits for young customers.
        - Offer higher credit lines to customers with “great balances.”
        - Apply tiered interest rates by combining age, duration, and account data.

For Modeling, we used a Random Forest Classifier for its interpretability and robust handling of mixed-type data.

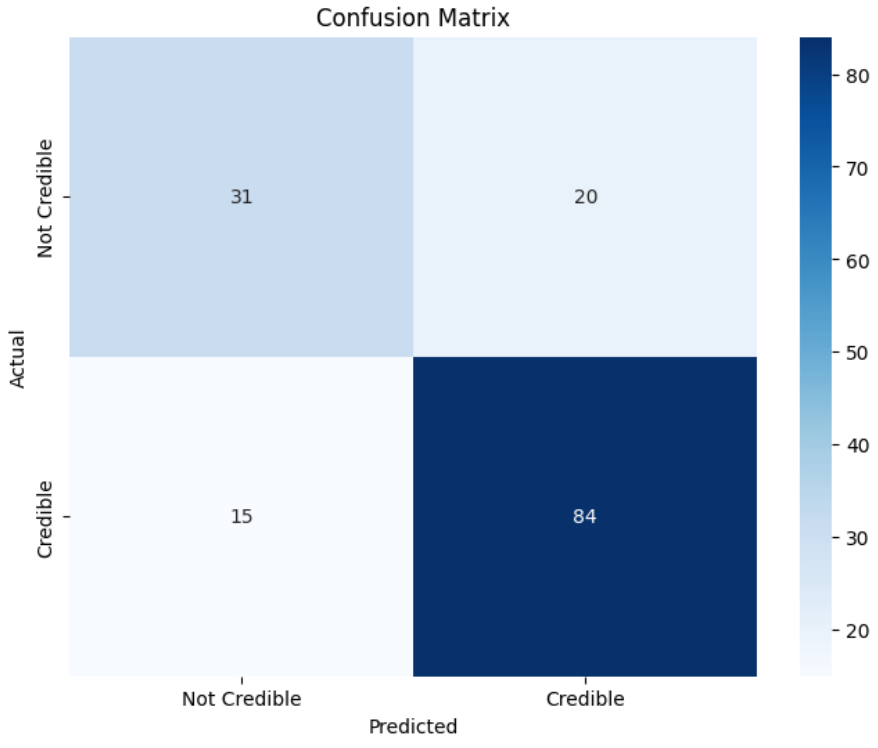
Model Features:

```
features = ['Age_years', 'Duration_of_Credit_monthly', 'Account_Balance',
'No_of_Credits_at_this_Bank', 'Sex_Marital_Status']
target = 'Creditability'
```

Model Performance (on test set):

- Accuracy: 77%
- Precision (class 1): 81%
- Recall (class 1): 85%
- F1-Score (class 1): 83%
- Macro Average F1-Score: 73%

This suggests the model is particularly strong at identifying credible customers (class 1), which is valuable for lending institutions.



Why Fintech Companies Use ML for Credit Risk

1. **Faster Decision Making**: Unlike traditional banks that require manual underwriting, ML models can analyze hundreds of variables in seconds.
2. **Higher Approval Rates with Controlled Risk**: By assessing non-traditional data (e.g., digital footprints, mobile usage), ML can identify low-risk individuals previously unscored by credit bureaus.
3. **Dynamic Credit Scoring**: Real-time updates to a user's score allow fintech lenders to adjust offers instantly, enabling personalized loans.
4. **Cost-Efficiency**: Automated ML pipelines reduce operational costs, allowing fintech companies to offer better interest rates and user experience.
5. **Geographic/Regional Adaptation**: ML models can adapt by region or city based on local economic conditions and repayment behavior.

Suggestions to Improve the Model Further for Real World Adoption

1. **Feature Engineering**:
    - Incorporating granular data such as loan purpose, employment type, housing situation, or utility payment history (as recommended by Thomas et al.) can greatly enhance model performance.
    - Time-sensitive data, like recent delinquencies or income variation, adds predictive power.
2. **Model Diversity**:
    - XGBoost, LightGBM, and TabNet offer better performance on structured data than traditional models.
    - Studies like Lundberg et al. (2020) highlight how explainability tools (e.g., SHAP) can make these models interpretable for financial regulators.
3. **Probability-Based Scoring System**:
    - Shift from binary classification to probability-based scores (0–1), similar to a credit score.
    - Risk bands:
        - .0.75: High Risk – Likely default – Reject or high interest rate
        - 0.45–0.75: Medium Risk – Manual review or restricted offers
        - < 0.45: Low Risk – Fast-track approvals.
    - Used by firms like Upstart and Zest AI.

Train a Classification Model That Outputs Probabilities

Most classification models can return probability estimates using `.predict_proba()` (in scikit-learn) or `.predict()` with `probability=True` in some cases. The classification models can be logistic regression, random forest classifier, XG Boost, etc.

Customer ID	Predicted Prob (Default)	Credit Score	Risk Band
C123	0.82	376	High Risk
C124	0.32	727	Low Risk

- Advantage of Probability scoring over Normal Classification
  1. Granular Risk Assessment
     - Binary: Says only "default" or "no default"
     - Probability: Says "this customer has a 72% chance of defaulting"
     - This allows you to differentiate between:
       - Someone with 0.51 probability (barely risky)
       - Someone with 0.91 probability (extremely risky)
         Even though both would be labelled "default" in binary classification.
- Dynamic Credit Limits: Allocate credit based on risk:
  - A low-risk applicant might get ₹1,00,000.
  - A borderline case gets ₹25,000–₹40,000.
  - Reduces lenders' capital exposure and defaults, especially when there are high approval rates.

4. **Behavioral Segmentation**:
   - Apply K-Means/Hierarchical Clustering to form borrower personas:
     - Cluster A: Young salaried professionals – moderate risk.
     - Cluster B: Seasoned customers with stable repayment, low risk.
     - Cluster C: Irregular earners or high credit appetite – high risk.
   - Benefits:
     - Personalize product recommendations
     - Adjust risk strategy dynamically
     - Train targeted models per segment for better performance
5. **Feedback Loop for Continuous Learning**:
   - Automate model retraining every quarter using updated repayment data.
   - Capture macroeconomic shifts (e.g., recession, employment changes).
   - Enables adaptive credit scoring.

6. **Final Recommendations for Real World Adoption**:
    - Explainability & Compliance: Ensure model decisions can be explained using tools like SHAP, LIME—important for regulatory compliance.
    - Ethical AI: Audit models for bias (gender, age, region) to prevent discrimination in lending.
    - Deployment Pipeline: Implement robust APIs, monitoring dashboards, and fallback logic in case the model confidence is low.
    - Data Partnerships: Collaborate with telecoms, e-commerce, or UPI apps to access more real-time behavioral signals.
    - User Education: Offer personalized tips to users to improve their credit standing—helps build trust and long-term customer loyalty.

Conclusion: This project highlights the power of machine learning in revolutionizing credit risk assessment. By leveraging accessible datasets and training robust classifiers like Random Forest, fintech companies can make informed, real-time credit decisions, enhancing inclusivity and driving financial innovation. The integration of account data, demographic features, and credit history allows for nuanced, data-driven lending with lower risk exposure.