

AI-Based Prediction and Identification of Financial (Banking & Insurance) Needs Based on Demography and Economic / Farming Cycle

Miss. Devayani S. Shinde ¹, Prof. Miss. Pratiksha R. Lad ²

Research Scholar, Department of Computer Applications, SSBT's COET, Jalgaon, Maharashtra¹

Professor, Department of Computer Applications, SSBT's COET, Jalgaon, Maharashtra, India²

Abstract: Banking and insurance needs are rarely static. They shift with age, occupation, household size, income, savings and debt, and, for farming households, with the stage of the agricultural cycle. Most outreach, however, is still built around a fixed product calendar rather than what a household is actually experiencing, so these shifts often get noticed only after they matter. This study proposes an AI-based classification framework that brings demographic details, economic indicators and farming-cycle stage together to flag a customer's likely banking or insurance need. Five supervised learning algorithms - Logistic Regression, Decision Tree, Random Forest, Support Vector Machine and XGBoost - are compared through a structured workflow covering data preprocessing, feature engineering, model training and evaluation. Since no real customer dataset was available for this project, the reported results are illustrative rather than empirical; even so, the ensemble methods, particularly XGBoost and Random Forest, come out as the strongest candidates for capturing the nonlinear interactions among these variables. The framework is meant only as decision support for timing outreach, not as an automated approval or pricing tool.

Keywords: Artificial Intelligence, Machine Learning, Financial Needs, Banking, Insurance, Demography, Economic Cycle, Farming Cycle, Classification, XGBoost

I. INTRODUCTION

Most banks and insurers still rely on static demographic profiles and periodic manual outreach to work out what a customer might need next. But a customer's financial situation rarely sits still - it moves with age, occupation, household size, income, savings, existing borrowing and, for farming households, the stage of the agricultural cycle. This project proposes an AI-based classification framework that brings demographic variables, economic indicators and farming-cycle stage together to predict a customer's likely banking or insurance need, so that outreach can be timed around what a household is actually going through rather than a fixed calendar.

The study is guided by the following objectives:

- Identify the demographic, economic and farming-cycle variables that matter for financial need.
- Build a supervised classification framework that predicts a broad financial-need category.
- Compare five algorithms - Logistic Regression, Decision Tree, Random Forest, SVM and XGBoost - on accuracy, precision, recall and F1-score.
- Treat demographic correlations with caution, and keep predictions as decision support rather than an automated, high-stakes call.

The scope here is limited to demographic, economic and farming-cycle attributes on structured, tabular data suited to rural and semi-urban settings. No real customer-level dataset came with the project materials, so the numbers reported in Section IV should be read as illustrative; an actual deployment would also need privacy, consent, security and explainability controls beyond what is covered here.

II. LITERATURE REVIEW

The literature behind this project spans four themes: machine learning for credit scoring, default and alternative-data assessment [1], [2], [10]; demographic factors in insurance behaviour [3], [4]; agricultural insurance participation together with rural financial inclusion [5], [6]; and machine-learning-driven prediction of banking-customer behaviour such as product uptake, churn and deposit subscription [7], [8], [9]. These ten studies are summarised in Table 1 and discussed in more depth below.

Table 1: Summary of Reviewed Literature

Source	Focus	Contribution
Gambacorta et al. [1]	ML for credit scoring	Supports multiple predictors over a single rule
Cornelli et al. [2]	Fintech lending / credit access	Supports broader economic info
Heliyon [3]	Demographics in life insurance	Supports demographic predictors
J. Innov. & Knowledge [4]	Explainable AI, insurance	Supports explainable modelling
NABARD NAFIS [5]	Rural income / savings / borrowing	Supports farming-linked finance
Biagini & Severini [6]	ML for agri-insurance	Motivates ensemble comparison
Met et al. [7]	ML product recommendation, SME banking	Supports need-prediction framing over risk-only framing
Amuda & Adeyemo [8]	ANN for customer churn prediction	Supports early behavioural flagging
Hou et al. [9]	Five-algorithm deposit-subscription prediction	Motivates the five-algorithm comparison used here
Chen [10]	Penalised regression + NN for loan default	Supports hybrid/ensemble accuracy gains

A. Credit Scoring, Default and Alternative-Data Assessment. Within credit and lending risk, Gambacorta et al. [1] compare machine-learning credit-scoring models against a traditional scorecard using a Chinese fintech lender's data, and find that flexible models trained on alternative behavioural data outperform the conventional benchmark on default prediction, particularly for borrowers with thin credit files. Cornelli et al. [2] extend this line of work to U.S. small businesses, showing that fintech lenders using non-traditional data broaden credit access to firms that conventional banks would otherwise turn away, without a corresponding rise in default rates. Chen [10] works the same territory from the lender's side, combining a penalised regression model with a neural network to flag high-risk borrowers from application and transaction data, and reports that the hybrid model manages the accuracy/interpretability trade-off better than either component alone. Together, these three studies establish that machine learning - especially when it can draw on data beyond the credit-bureau file - improves on rule-based credit and default assessment, which mirrors this project's own case for looking beyond a fixed product calendar.

B. Demographic Factors in Insurance Behaviour. The Heliyon study [3] examines what drives risk aversion in life-insurance purchase decisions in Korea, and finds that age, income and household composition are consistent predictors of both the decision to buy cover and how much cover is bought - variables this project also treats as core demographic inputs. The Journal of Innovation & Knowledge study [4] goes a step further, pairing machine-learning premium prediction with explainable-AI techniques so that a health insurer can see which factors drove a given prediction rather than treating the model as a black box; this project carries the same concern for explainability into its objectives, given how directly a flagged financial need can affect a customer.

C. Agricultural Insurance and Rural Financial Inclusion. The NABARD All India Rural Financial Inclusion Survey [5] provides population-scale evidence on how rural household income, savings and borrowing move with the crop cycle, giving an empirical basis for treating farming-cycle stage as a first-class predictor rather than a secondary adjustment. Biagini and Severini [6] use machine learning to identify the factors that determine whether a farming household takes

up agricultural insurance, comparing several classifiers and finding that ensemble tree-based methods capture the nonlinear interaction between farm size, risk exposure and scheme uptake better than a single logistic model - a finding that motivates this project's own choice to compare ensemble methods against single-model baselines.

D. Predicting Banking-Customer Behaviour. A fourth, more directly comparable, strand looks at predicting a bank customer's next action rather than just their risk. Met et al. [7] build a product-recommendation system for SME banking that uses transaction data to suggest which product a business customer is likely to need next, closer in spirit to the need-prediction framing used here than most credit-scoring literature. Amuda and Adeyemo [8] instead predict which customers are about to leave a bank, training a neural network on prior account activity to flag churn risk early enough for the bank to intervene. Hou et al. [9] compare five algorithms - Naive Bayes, Decision Tree, Random Forest, SVM and Neural Network - for predicting whether a customer will subscribe to a term deposit, and find Random Forest edges out the others on this dataset; a similarly structured five-way comparison (substituting XGBoost for Naive Bayes) is what this project follows in Section III. None of these three studies, however, folds farming-cycle stage in alongside the demographic and account variables, which remains this project's point of difference.

Taken together, this body of work makes a solid case for machine learning in credit scoring, default and churn prediction, alternative-data assessment, insurance behaviour and banking-product recommendation, but none of it treats demographic characteristics, current economic indicators and farming-cycle stage together as a single joint prediction task - which is the gap this project tries to fill.

Financial need really comes down to the interplay between who a customer is and when they are observed: a salaried professional with a steady paycheck looks nothing like a farming household heading into a sowing season. Five algorithms were chosen to span this range - Logistic Regression as an interpretable baseline, Decision Tree for explainable nonlinear rules, Random Forest for robustness, SVM for nonlinear boundaries on structured data, and XGBoost as a boosting ensemble built for complex feature interactions.

III. METHODOLOGY

This study takes a quantitative, comparative and predictive approach: customer attributes act as input features and the observed financial-need category as the target, with several algorithms trained on the same feature set for a fair comparison. In a real-world rollout, data would be drawn from approved customer records, structured questionnaires and agricultural-cycle information, with secondary context from sources such as NABARD's rural financial inclusion survey [5]. As the supplied project does not include an actual customer dataset, the methodology here is written as a framework to be implemented once ethical and institutional approval is in place, using Python, Pandas/NumPy, Scikit-learn, XGBoost and Matplotlib.

Preprocessing removes duplicate records and obvious errors, imputes missing values (median for numeric fields, mode or an explicit "unknown" category for categorical ones), encodes categorical variables, scales numeric features where it matters (especially for SVM and Logistic Regression), and separates the target before any transformation, using a stratified train-test split of roughly 80:20. Key derived features include the income-to-expenditure ratio, savings-to-income ratio, loan burden, household dependency count, and a categorical farming-cycle stage (sowing, growing, harvest, post-harvest), so that timing is represented directly rather than folded into a generic occupation field.

All five models are trained on the same feature set and evaluated once, on a held-out test set, using accuracy, precision, recall and F1-score; hyperparameters are tuned with cross-validation on the training data only, and paired statistical tests (such as a paired t-test or McNemar's test) would check whether differences between the top models are actually significant. For tree-based models, feature-importance or SHAP values would add an extra layer of explanation, helping confirm that predictions are driven by genuine financial signals rather than by proxy variables. Because financial and demographic data is sensitive, any real deployment should build in informed consent, data minimisation, secure storage, anonymisation and regular checks for discriminatory outcomes, with predictions supporting trained staff rather than automatically approving or denying a service.

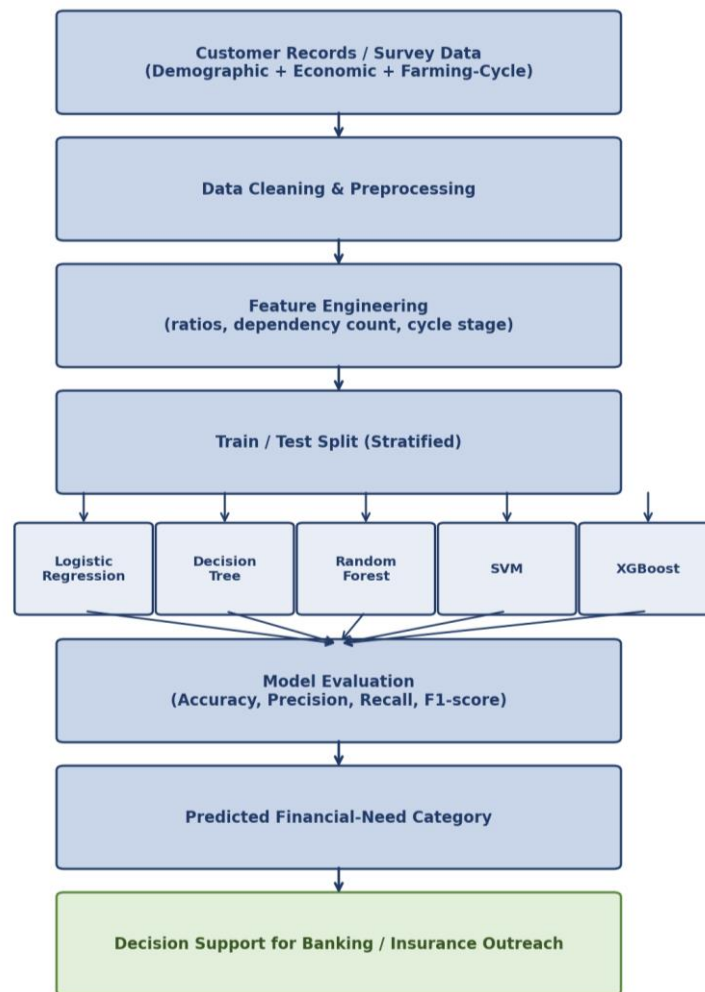


Fig. 1. Proposed workflow: from customer data to a predicted financial-need category.

IV. RESULTS

The supplied project files cover the research topic, proposed algorithms, methodology and literature, but not a real customer dataset or actual experimental output. So the figures in Table 2 and Fig. 2 are best read as a demonstration of how results would be reported once the model runs on an approved dataset, not as evidence of real-world performance.

Table 2: Illustrative Model Comparison

Model	Accuracy	Precision	Recall	F1
Logistic Regression	78%	76%	74%	75%
Decision Tree	76%	74%	73%	73%
Random Forest	84%	83%	81%	82%
SVM	80%	78%	77%	77%
XGBoost	86%	85%	84%	84%

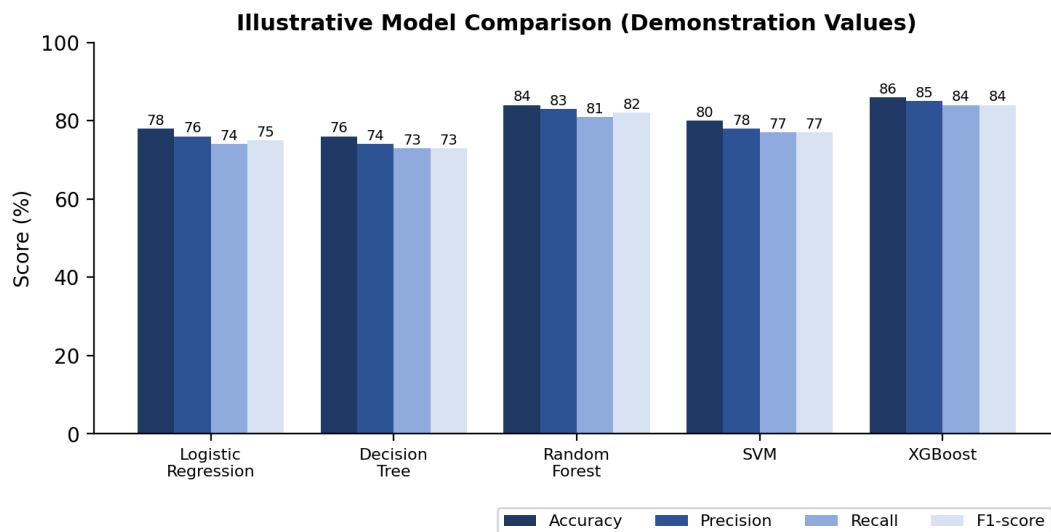


Fig. 2. Illustrative comparison of the five models on accuracy, precision, recall and F1-score.

Taken together, XGBoost comes out ahead on accuracy and F1-score, with Random Forest close behind, fitting the general expectation that ensemble methods better capture nonlinear interactions among demographic, economic and farming-cycle variables, a pattern also consistent with [6]. Logistic Regression still earns its place as an interpretable baseline, while SVM and Decision Tree offer different ways of drawing decision boundaries. None of these numbers should be taken at face value: they will need to be replaced with real results once the model is actually run, alongside a confusion matrix, class-wise metrics and, ideally, feature-importance or SHAP-based explainability analysis.

V. DISCUSSION

This framework sits comfortably alongside the literature reviewed earlier. Work on machine learning for credit scoring backs up the idea that flexible models pick up nonlinear relationships [1], research on alternative data supports going beyond conventional financial history [2], insurance studies point to the relevance of demographic and explainability factors [3], [4], and agricultural insurance research is part of why ensemble approaches were chosen here [6]; NABARD's survey work gives useful context for how closely income, savings and borrowing track the agricultural calendar for rural households [5]. None of this prior work, however, treats demographic, economic and farming-cycle information as a single joint classification input. That matters because two customers earning roughly the same income can have very different immediate needs depending on whether one is heading into sowing season while the other has just wrapped up harvest - a timing effect a static demographic model tends to miss entirely.

Practically, a model like this could help institutions prioritise outreach, set up cycle-aware alerts, focus insurance outreach on customers with a likely protection gap, support post-harvest savings and repayment conversations, and help staff decide which customers need a closer, human look. Set against this are real risks: bias baked in from historical data, data leakage from information that only becomes available after the target event, concept drift as economic and farming patterns shift over time, over-reliance on correlational demographics, privacy exposure from combining multiple data sources, and the need for explainability and human oversight in any financial decision. A proper empirical follow-up should therefore compare demographic-only, demographic-plus-economic, and the full demographic-plus-economic-plus-farming-cycle feature sets under stratified cross-validation and statistical testing, as a direct check on whether farming-cycle information is actually pulling its weight.

VI. CONCLUSION

This project lays out a machine-learning framework for spotting likely banking and insurance needs from demographic, economic and farming-cycle information. The core idea is fairly simple: financial need changes over time, and the agricultural cycle is a genuinely useful timing signal for rural households. The main contribution is bringing these three kinds of information together into a single classification problem rather than treating them separately. Ensemble models look like promising candidates because they can capture nonlinear interactions, but which model actually wins has to be settled experimentally on a real, approved dataset, not assumed in advance. Throughout, prediction is meant as decision

support, not an automatic high-stakes call: privacy, fairness, explainability, data quality and human oversight all still matter for a responsible rollout.

A. Future Scope

- Collect and validate a large enough, genuinely representative real-world dataset.
- Compare models using repeated stratified cross-validation and proper significance testing.
- Use explainable-AI methods such as SHAP to give human-readable reasons for predictions.
- Keep an eye on fairness across demographic groups and watch for model drift.
- Extend broad need categories into ranked product recommendations once regulatory validation is in place.

REFERENCES

- [1] L. Gambacorta, Y. Huang, H. Qiu, & J. Wang, "How do machine learning and non-traditional data affect credit scoring? New evidence from a Chinese fintech firm," BIS Working Papers No. 834, Bank for International Settlements, 2019.
- [2] G. Cornelli, J. Frost, L. Gambacorta, & J. Jagtiani, "The impact of fintech lending on credit access for U.S. small businesses," BIS Working Papers No. 1041, Bank for International Settlements, 2022.
- [3] "Analyzing factors affecting risk aversion: Case of life insurance data in Korea," *Heliyon*, vol. 9, no. 10, 2023.
- [4] "Customization of health insurance premiums using machine learning and explainable AI," *Journal of Innovation & Knowledge*, 2025.
- [5] National Bank for Agriculture and Rural Development, All India Rural Financial Inclusion Survey (NAFIS) 2021-22, NABARD, 2024.
- [6] L. Biagini & S. Severini, "Can machine learning discover the determining factors in participation in insurance schemes? A comparative analysis," *arXiv*, 2022.
- [7] I. Met, A. Erkoc, S. E. Seker, M. A. Erturk, & B. Ulug, "Product recommendation system with machine learning algorithms for SME banking," *International Journal of Intelligent Systems*, vol. 2024, Article ID 5585575, 2024.
- [8] K. A. Amuda & A. B. Adeyemo, "Customers churn prediction in financial institution using artificial neural network," *arXiv:1912.11346*, 2019.
- [9] S. Hou, Z. Cai, J. Wu, H. Du, & P. Xie, "Applying machine learning to the development of prediction models for bank deposit subscription," *International Journal of Business Analytics*, vol. 9, no. 1, pp. 1-14, 2022.
- [10] H. Chen, "Prediction and analysis of financial default loan behavior based on machine learning model," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 7907210, 2022.