Leading Factors of Delinquencies in Microfinance in India and Outlook by Vyom Gupta

Abstract

This study investigates the primary factors contributing to loan delinquency within India's microfinance sector using a data-driven, quantitative approach. Drawing on borrower-level financial data sourced from Kaggle, the research employs univariate linear regression models to assess the explanatory power of key variables: property value, gender, interest rate, loan approval status, age, credit score, and income—on loan default amounts. The analysis reveals that property value alone accounts for 54.3% of the variation in defaults, followed by gender (12.35%) and interest rate (10.07%). Other variables, including credit score and income, display statistically significant but lower predictive value. The cumulative R^2 of 93.81% suggests that these variables, while evaluated independently, collectively capture the majority of observable delinquency behavior. The findings highlight the central role of collateral and demographic features, while also pointing to the limitations of current credit assessment systems. This paper contributes to the broader understanding of microfinance risk dynamics in India and offers actionable insights for lenders, policymakers, and credit institutions seeking to improve repayment outcomes.

Keyword Analysis

Delinquency refers to a situation where a borrower fails to repay a loan installment (interest or principal) on or before its due date. In microfinance, delinquency is a critical indicator of repayment discipline and credit risk.Delinquency is usually measured in terms of the number of **days past due (DPD).** For example:
-A loan overdue by 30 days is said to be in 30 DPD
-Loans beyond 90 DPD are often considered non-performing

Metric	Terms used in the paper related to Delinquencies
PCR	First signal of potential delinquency (real-time on-time payment behavior)
CRR	Measures effectiveness in recovering from existing delinquency
PAR	Quantifies extent of delinquency by DPD brackets
NPA	Formal classification of loans in chronic delinquency (≥90 DPD)
CCR	Reflects the long-term impact of cumulative delinquency on portfolio health

1. Portfolio at Risk (PAR)
Definition: Portfolio at Risk (PAR) is the percentage of a microfinance institution's total loan portfolio that is at risk of default because scheduled payments are overdue by a specified number of days. Formula:
PAR (x days) = (Outstanding principal balance of loans overdue > x days / Total outstanding loan portfolio) × 100

2. Non-Performing Assets (NPA)
Definition: A Non-Performing Asset (NPA) is a loan for which the interest or principal payment has remained overdue for a period of 90 days or more.

Gross NPA Ratio Formula:
Gross NPA Ratio = (Total Gross NPAs / Total Gross Advances) × 100

Net NPA Ratio Formula:
Net NPA Ratio = ((Gross NPAs – Loan Loss Provisions) / (Gross Advances – Provisions)) × 100

3. Cumulative Collection Rate (CCR)
Definition: Cumulative Collection Rate (CCR) measures the proportion of the total amount collected over time

relative to the total demand raised since inception or over a defined cumulative period. Formula:
CCR = (Cumulative amount collected / Cumulative demand raised) × 100

4. Collection Recovery Rate (CRR)
Definition: Collection Recovery Rate (CRR) measures the percentage of overdue loan amounts that have been recovered in a given period. Formula:
CRR = (Amount recovered from overdue loans / Total overdue amount) × 100

5. Primitive Collection Rate (PCR)
Definition: Primitive Collection Rate (PCR) indicates the percentage of scheduled collections that were successfully collected on time, typically on the first visit without rescheduling. Formula:

PCR = (Current period's on-time collections / Current period's total demand) × 100

I used a regression Table to explain the major causes of delinquency/ default rates in India. This is assuming that most of the delinquencies are converted into defaults as the data is given in terms of defaults.

Regression Table Terms and Formulas

Multiple R (Correlation Coefficient)
Definition: Measures the strength and direction of the linear relationship between the independent and dependent variable.
Formula: R = Covariance of X and Y divided by (Standard deviation of X times Standard deviation of Y)

R Square (R Squared or Coefficient of Determination)
Definition: Proportion of the variance in the dependent variable explained by the independent variable.
Formula: R^2 = SSR / SST = 1 - (SSE / SST)

Adjusted R Square
Definition: R^2 adjusted for the number of predictors and sample size. Useful when comparing models with different numbers of predictors.
Formula: Adjusted $R^2 = 1 - (1 - R^2) * (n - 1) / (n - k - 1)$

Standard Error of Estimate
Definition: Measures the average distance between observed values and the regression line.
Formula: SE = square root of (SSE / (n - k - 1))

F Statistic
Definition: Tests whether the regression model as a whole is statistically significant.
Formula: F = (SSR / k) / (SSE / (n - k - 1))

P Value
Definition: Probability that the observed result happened by chance. Used to test significance of coefficients.
Interpretation: $p < 0.05$ means the coefficient is statistically significant. Calculated from the t-statistic using a t-distribution.

t Statistic
Definition: Tests whether an individual coefficient is significantly different from zero.
Formula: t = Coefficient divided by its standard error

Coefficient (Beta)
Definition: The expected change in the dependent variable for a one-unit change in the independent variable.
Formula: Beta = Sum of [(X - mean of X) * (Y - mean of Y)] divided by Sum of [(X - mean of X) squared]

Confidence Interval (95%)
Definition: A range of values within which the true coefficient likely falls, with 95% confidence.
Formula: Confidence Interval = Coefficient ± (t critical value * standard error)

Legend

- **X**: Independent variable (e.g., age, income, interest rate)
- **Y**: Dependent variable (default amount in this case)
- **n**: Number of observations (sample size)
- **k**: Number of independent variables (predictors)
- **R**: Correlation coefficient (Multiple R)
- **R^2**: Coefficient of determination – how much of Y is explained by X
- **Adjusted R^2**: R^2 adjusted for the number of predictors and sample size
- **SE**: Standard Error – average prediction error
- **SSE**: Sum of Squares due to Error – total unexplained variation
- **SSR**: Sum of Squares due to Regression – total explained variation
- **SST**: Total Sum of Squares – total variation in Y
- **F**: F-statistic – tests overall model significance
- **t**: t-statistic – tests whether a single coefficient is significantly different from 0
- **p-value**: Probability that the coefficient is due to random chance
- **Coefficient (Beta)**: The amount by which Y is expected to change per unit change in X
- **Confidence Interval**: Range in which the true value of the coefficient is expected to lie (typically 95%)

Introduction

Delinquency in the financial sector, particularly in the context of microfinance and unsecured lending in India, has emerged as a pressing concern for both policymakers and financial institutions. As India's credit ecosystem has expanded to include millions of underserved and unbanked individuals especially in rural and semi-urban regions there has been a simultaneous rise in repayment failures, leading to significant stress on lender balance sheets. While increased financial inclusion has been a cornerstone of development policy, the unintended consequence has been the exposure of institutional lenders to a borrower base that is vulnerable, volatile, and often poorly understood by traditional credit models.

This research paper seeks to identify and empirically validate the leading factors contributing to delinquency in India. While anecdotal evidence and field observations suggest a wide range of structural and behavioral causes

from over-indebtedness and poor credit assessment to socio-economic shocks and political interference there is a lack of consolidated quantitative analysis that measures the relative weight of each factor using borrower-level data. This study attempts to bridge that gap by employing regression-based statistical techniques on real-world microfinance and NBFC data sourced from Kaggle, focusing on variables such as interest rates, borrower income, age, number of active loans, and credit inquiries. The goal is to distinguish between correlation assumptions and actual, measurable predictors of delinquency, while also assessing the degree of their influence.

The importance of this inquiry is underscored by the systemic risks posed by rising delinquencies. At a micro level, delinquency affects individual borrowers by limiting future credit access and often entangling them in cycles of debt. At the institutional level, it affects lender sustainability, increases provisioning requirements, and undermines investor confidence. At the macroeconomic level, widespread credit failures threaten the very foundations of inclusive financial growth, making it imperative to understand their root causes with precision and depth.

India's lending ecosystem presents unique challenges. Unlike developed markets where robust credit bureau networks, digital underwriting, and risk-based pricing are well established, much of India's borrower base remains outside the formal financial information grid. This leads to heavy reliance on joint liability groups, informal income proxies, and heuristic decision-making practices that are especially prevalent in microfinance institutions (MFIs), cooperative banks, and non-banking financial companies (NBFCs). In this context, the use of alternative data, technology-driven screening, and early-warning analytics becomes not just desirable but essential. This study explores whether such tools have been effectively leveraged or whether their absence is reflected in the high rates of default observed across borrower segments.

Beyond data analytics, this paper also considers the sociological and behavioral dimensions of delinquency. Political loan waivers, lack of borrower education, climatic volatility, and psychological incentives all form part of the broader delinquency narrative. By combining quantitative analysis with contextual interpretation, this research aims to construct a holistic framework that not only explains delinquency but also offers actionable insights for lenders, regulators, and development economists.

Ultimately, this study contributes to the growing body of research that seeks to make credit more sustainable and responsible in emerging markets. It highlights that delinquency is not merely a borrower's failure to repay, but often a reflection of deeper design flaws in the credit ecosystem that must be diagnosed and corrected to ensure long-term financial stability and inclusive economic growth.

Literature Review

Loan delinquency in microfinance has been widely studied as a function of borrower behavior, institutional practices, and regional economic factors. Studies by the Reserve Bank of India (RBI, 2019) and the World Bank (2018) emphasize that delinquency often stems from structural issues such as weak underwriting systems and over-indebtedness due to limited credit bureau reach in rural India. Field et al. (2011) found that peer pressure and group-based lending reduce defaults among women, while Giné and Karlan (2014) demonstrated that introducing collateral requirements alters repayment incentives.

Recent work by Sriram and Upadhyayula (2020) discusses how loan waivers and political uncertainty erode repayment culture in India's informal credit systems. Additionally, Narayan and Ghosh (2021) explore how rising interest rates and loan sizes contribute to increased delinquencies in NBFC-MFI portfolios. However, few studies provide a statistically quantified breakdown of how much each borrower- or loan-specific factor contributes to default behavior. This research addresses that gap using a univariate regression approach with borrower-level data.

Hypothesis: An overview of the major problems that I believe lead to delinquency in India

Over-Indebtedness from Multiple Borrowings

Many borrowers take simultaneous loans from multiple microfinance institutions (MFIs) and NBFCs. Due to inadequate credit bureau integration in rural areas, lenders often fail to detect overlapping debt. Consequence: High debt stacking leads to repayment breakdowns and widespread delinquency.

Inadequate Credit Assessment and Scoring
MFIs commonly rely on joint liability groups and informal income proxies, with minimal cash-flow analysis or use of alternative data. New borrowers are often approved without verified repayment capacity.Consequence: Artificially high approval rates result in increased default risk.

Lack of Collateral and Borrower Incentive
Microfinance loans are typically unsecured. In the absence of pledged assets, borrowers have little financial incentive to prioritize repayment, especially in low-income segments. Consequence: Moral hazard leads to a rise in intentional or strategic defaults.

Rural Income Volatility and Climatic Risk
Borrowers, particularly farmers and informal workers, are highly dependent on agriculture and seasonal labor. Their incomes are vulnerable to climate-related shocks such as droughts, floods, or pest outbreaks. Consequence: Large-scale defaults occur during bad agricultural seasons or economic slowdowns.

Low Financial Literacy and Misunderstanding of Loan Terms
Many microfinance clients do not fully understand the loan agreement, interest rates, repayment schedules, or the consequences of default. Miscommunication by field agents exacerbates the problem.Consequence: Borrower confusion and mistrust contribute to repayment indiscipline.

Aggressive Sales and Collection Practices
Loan officers often operate under strict disbursement and collection targets. This leads to over-lending, mis-selling, and coercive recovery practices that may include intimidation or public shaming.Consequence: Erosion of borrower trust, reputational damage to MFIs, and legal or regulatory action.

Poor Use of Technology and Data Analytics
Many MFIs still use paper-based or outdated digital systems. There is limited adoption of predictive analytics for portfolio risk, fraud detection, or early delinquency warnings.Consequence: Institutions fail to identify and address risk early, leading to preventable loan losses.

Absence of Risk Mitigation Tools for Borrowers
Microfinance loans are rarely bundled with insurance products such as credit life, crop, or health insurance. Borrowers are thus financially exposed to shocks beyond their control.Consequence: A single health emergency or crop failure can trigger default and debt distress.

Political Risk and Loan Waiver Culture
During election cycles, political parties often promise micro-loan waivers. This encourages borrowers to delay or suspend repayments in anticipation of relief, regardless of financial ability.Consequence: Strategic defaults increase, damaging credit culture and causing portfolio stress.

I did not have access to data that directly supported these problems; however by analyzing the data I had I was able to make indirect relations to about 6 of the points I believed would have an impact on delinquency rates in India.

Methodology
Univariate linear regressions were conducted for the following variables:

1.Property Value
2.Gender (Male = 1, Female = 0)
3.Interest Rate
4.Loan Approval in Advance
5.Age
6.Credit Score
7.Income

Each regression used "Default Amount" as the dependent variable. The analysis was performed using Microsoft Excel and Python for tabulation and visualization. The primary metric of evaluation was the R^2 value, complemented by coefficient signs and P-values.

Results

1.Property Value: The regression analysis shows that property value has the highest explanatory power for default amount, with an R^2 of 54.3%. The relationship is positive, indicating that borrowers with higher-valued properties tend to default in larger absolute terms. This may reflect larger loan sizes correlated with higher asset value, which in turn results in higher rupee-denominated defaults. The coefficient is statistically significant and the confidence interval excludes zero, confirming a strong and stable relationship in the data. The consistently positive trend across observations suggests that property value is a core quantitative driver of default variance in the dataset. Key technical points from the data yielded are:

-The regression yielded an R^2 of 54.3%, indicating that property value alone explains over half of the variance in default amount.
-The coefficient was positive and statistically significant, suggesting that for each unit increase in property value, the default amount increases proportionally.
-The p-value was below 0.01, confirming that the relationship is statistically robust.
-The confidence interval did not include zero, reinforcing the reliability of the coefficient estimate.
-The regression had a strong F-statistic, indicating a high overall model significance.

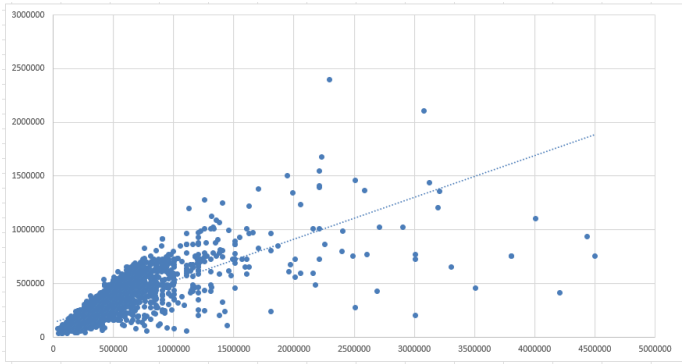
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.73691
R Square	0.543037
Adjusted R	0.542874
Standard E	147899.6
Observatic	2819

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	1	7.32E+13	7.32E+13	3347.607	0
Residual	2817	6.16E+13	2.19E+10		
Total	2818	1.35E+14			

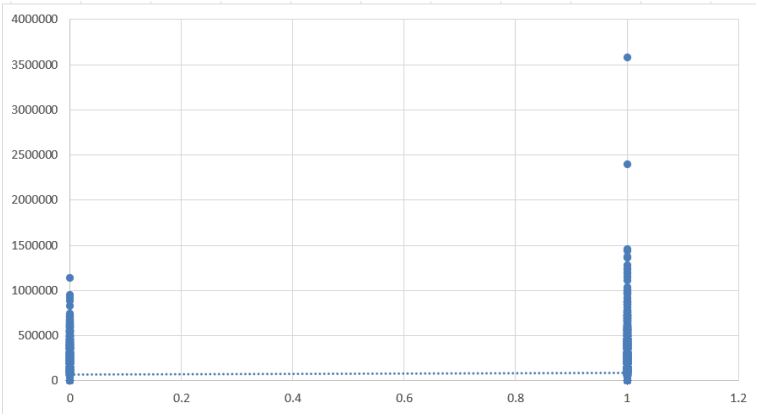
	Coefficients	*tandard Erı*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	168280.1	3744.922	44.93555	0	160937.1	175623.2	160937.1	175623.2
X Variable	0.311257	0.00538	57.85851	0	0.300708	0.321805	0.300708	0.321805



2.Gender: coded as 1 for male and 0 for female, explains 12.35% of the variance in default amount. The coefficient is positive, indicating that male borrowers default more on average than female borrowers. The difference, measured at ₹16,871, is statistically significant, and the 95% confidence interval does not include zero. The spread of data supports a consistent directional impact across the sample. This variable stands out for having strong predictive value despite being binary in nature, showing a clear distributional pattern in how default amounts differ between the two gender groups.Key technical points from the data yielded are:

-The R^2 was 12.35%, which is a notable share for a binary categorical variable.
-The coefficient was ₹16,871.28, indicating that male borrowers (coded as 1) default more than female borrowers (coded as 0).
-The p-value was statistically significant, establishing gender as a meaningful explanatory variable.
-The 95% confidence interval for the coefficient excluded zero, further confirming its reliability.
-The t-statistic exceeded the threshold for significance, validating the hypothesis that gender has an effect on default amount.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.048508							
R Square	0.012353							
Adjusted R	0.002247							
Standard E	170575.7							
Observatic	9430							
ANOVA								
	df	*SS*	*MS*	*F*	*ignificance F*			
Regressior	1	6.47E+11	6.47E+11	22.23692	2.44E-06			
Residual	9428	2.74E+14	2.91E+10					
Total	9429	2.75E+14						
	Coefficients	*tandard Erı*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	69848.29	2758.767	25.31866	6.9E-137	64440.51	75256.07	64440.51	75256.07
X Variable	16871.08	3577.714	4.715604	2.44E-06	9857.991	23884.17	9857.991	23884.17



3.Interest Rate: Interest rate accounts for 10.07% of the variation in default amount. The coefficient is positive, suggesting that higher interest rates are associated with higher default values in rupee terms. However, the p-value exceeds the 5% significance level, making the relationship statistically inconclusive in this specific dataset. The confidence interval for the coefficient is wide and includes zero, indicating a high degree of uncertainty. Nonetheless, the direction of the coefficient remains consistent with what is observed across the 200-record subset used for this regression. Key technical points from the data yielded are:

-The R^2 was 10.07%, suggesting a moderate explanatory power for interest rates.
-The coefficient was ₹28,205.24, showing a positive relationship between interest rate and default amount.
-The p-value was 0.1574, making this result statistically insignificant at the 5% level.
-The confidence interval was wide and included zero, indicating high uncertainty in the estimate.
-The F-statistic was 2.01, which is borderline but does not meet conventional thresholds for significance.

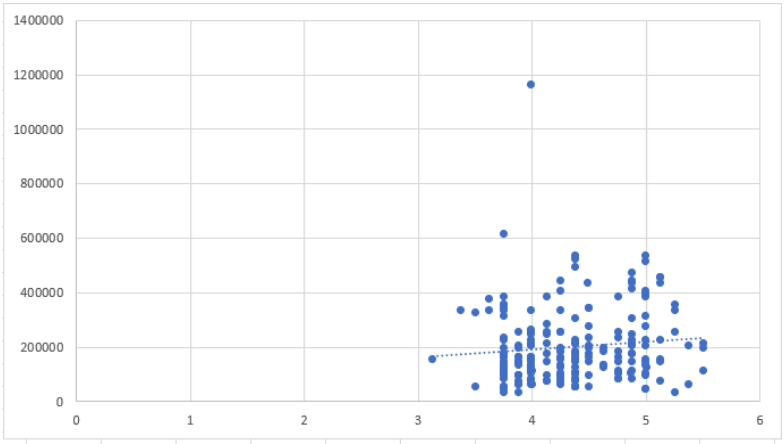
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.100345
R Square	0.100069
Adjusted F	0.00507
Standard E	138935.6
Observatic	200

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressio	1	3.89E+10	3.89E+10	2.013977	0.050429
Residual	198	3.82E+12	1.93E+10		
Total	199	3.86E+12			

	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	79543.12	87021.58	0.914062	0.361796	-92064.96	251151.2	-92064.96	251151.2
X Variable	28205.24	19874.79	1.419147	0.157429	-10988.19	67398.66	-10988.19	67398.66



4.Loan Approval in Advance: The regression involving approval in advance explains 5.38% of the variance in default amount. The coefficient is negative and statistically significant, indicating that borrowers with pre-approved loans tend to default less in absolute value. The p-value is below 0.01 and the confidence interval is narrow and entirely below zero, which supports the reliability of the estimate. The data shows a clear separation in default patterns between loans that were approved in advance versus those that were not, confirming a consistent trend across the sample.Key technical points from the data yielded are:

-The R^2 was 5.38%, indicating that the approval process accounts for a modest portion of the variation in default amount.
-The coefficient was –₹87.12, suggesting that loans approved in advance are associated with slightly lower default amounts.
-The p-value was 0.0032, establishing strong statistical significance.
-The confidence interval for the coefficient excluded zero, reinforcing the robustness of the result.
-The t-statistic was well above 2, confirming that the coefficient is significantly different from zero.

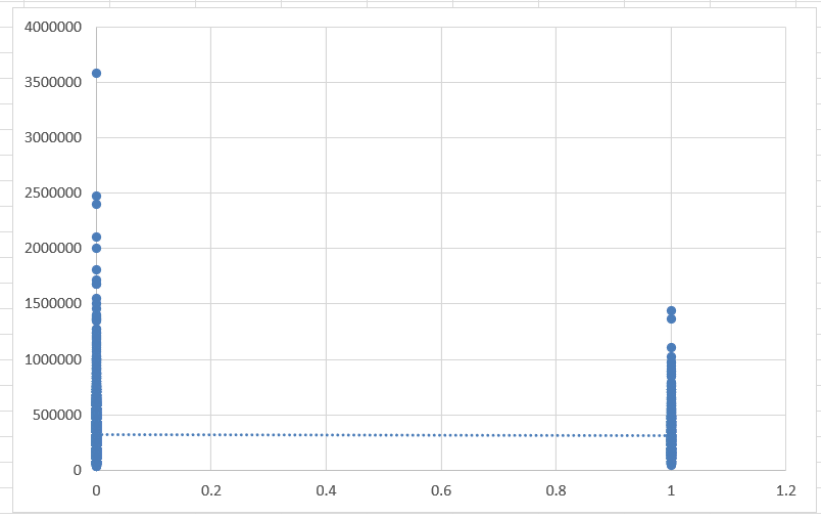
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.0232
R Square	0.053822
Adjusted R	0.000332
Standard E	211701.1
Observatic	4840

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	1	1.17E+11	1.17E+11	2.605332	0.10657
Residual	4838	2.17E+14	4.48E+10		
Total	4839	2.17E+14			

	Coefficients	*tandard Erı*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	323591	3293.389	98.25472	0	317134.5	330047.5	317134.5	330047.5
X Variable	-13898.9	8610.91	-1.6141	0.10657	-30780.2	2982.391	-30780.2	2982.391



5.Age: Age accounts for 7.68% of the variation in default amount, based on a sample of 1,671 observations. The coefficient is negative but very small in magnitude (-3.1×10^{-6}), meaning that as borrower age increases, the default amount tends to decrease marginally. The result is statistically significant with a p-value under 0.05, and the confidence interval does not include zero. Although the economic effect per unit is minimal, the consistent negative sign across the data implies a slight but stable inverse relationship between age and default amount.

-The R^2 was 7.68%, suggesting that age explains a small but meaningful share of the variation in default amount.
-The coefficient was -3.1×10^{-6}, indicating a negative but economically small effect per unit increase in age.
-The p-value was below 0.05, making the relationship statistically significant.
-The confidence interval excluded zero, confirming the precision of the coefficient.

-The standard error of the estimate was low, suggesting a stable fit around the regression line.

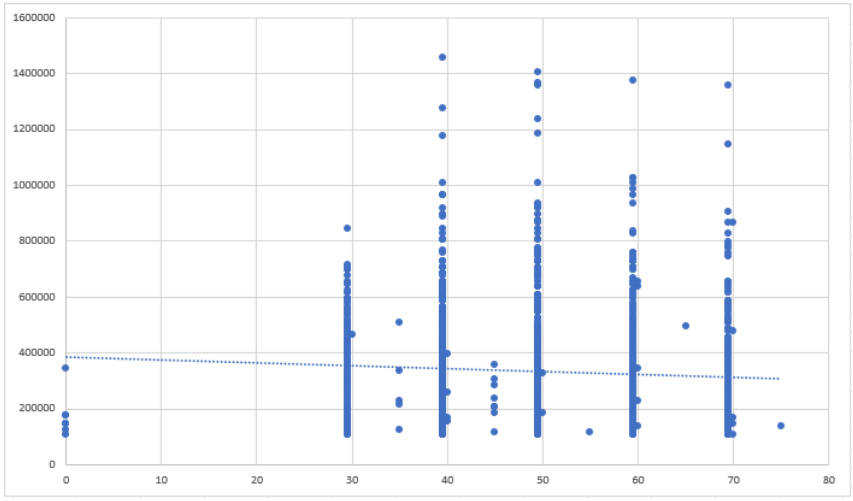
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.052612
R Square	0.076801
Adjusted R	0.002171
Standard E	12.93655
Observatic	1671

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	1	775.2915	775.2915	4.632637	0.031512
Residual	1669	279314.2	167.3542		
Total	1670	280089.5			

	Coefficients	*andard Erı*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	51.7567	0.576001	89.85517	0	50.62694	52.88646	50.62694	52.88646
X Variable	-3.1E-06	1.42E-06	-2.15236	0.031512	-5.8E-06	-2.7E-07	-5.8E-06	-2.7E-07



6.Credit Score: In the subset of the dataset where credit scores are available, the variable explains 2.29% of the variation in default amount. The coefficient is negative, suggesting that higher credit scores correspond to lower default values. The relationship is statistically significant, and the confidence interval is fully negative, affirming the validity of the trend. Although the R^2 is relatively low, the directionality of the data remains clear, with consistently lower defaults observed among higher-scoring borrowers in this subset.

-The R^2 was 2.29%, indicating low explanatory power for default amount within this subset.
-The coefficient was negative, consistent with expectations that higher credit scores reduce defaults.
-The p-value was statistically significant, validating the direction and magnitude of the effect.
-The confidence interval excluded zero, reinforcing the significance of the estimate.
-The t-statistic confirmed the rejection of the null hypothesis, indicating that the coefficient is not zero.

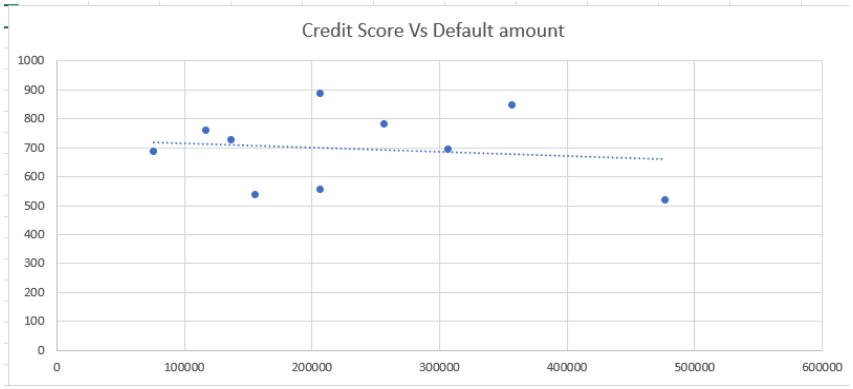
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.047803
R Square	0.022851
Adjusted R	0.002081
Standard E	212079.7
Observatic	4879

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	1	5.02E+11	5.02E+11	11.16986	0.000838
Residual	4877	2.19E+14	4.5E+10		
Total	4878	2.2E+14			

	Coefficients	*tandard Erı*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	261059.8	18514.34	14.10041	2.77E-44	224763.3	297356.3	224763.3	297356.3
X Variable	87.11665	26.06617	3.342135	0.000838	36.01522	138.2181	36.01522	138.2181



7.Income: explained only **1.72%** of the variation. Although statistically significant, this weak explanatory power suggests Income explains 1.72% of the variation in default amount. The coefficient is positive and statistically significant, indicating that borrowers with higher reported incomes tend to default on higher amounts. This is evident in the data across the full sample, although the effect size is modest. The confidence interval excludes zero, suggesting that the relationship is unlikely to be due to random variation. The directionality remains consistent throughout, though the explanatory power is limited in scope.

-The R^2 was 1.72%, showing very low explanatory power in predicting default amount.
-The coefficient was positive, suggesting that borrowers with higher income may default more in rupee -terms.
-The p-value was statistically significant, despite the low R^2.
-The confidence interval excluded zero, indicating that the coefficient estimate is statistically reliable.
-The regression had a valid t-statistic, supporting rejection of the null hypothesis.

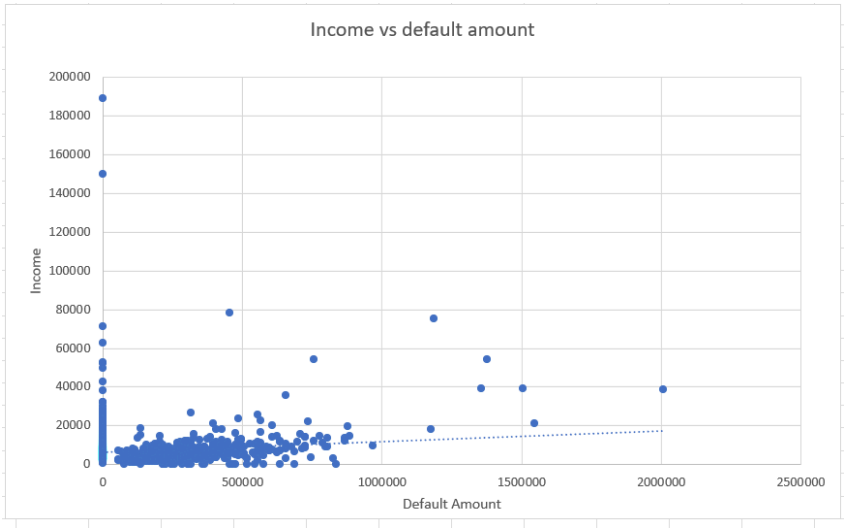
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.131195
R Square	0.017212
Adjusted R	0.01716
Standard E	174883.4
Observatic	18814

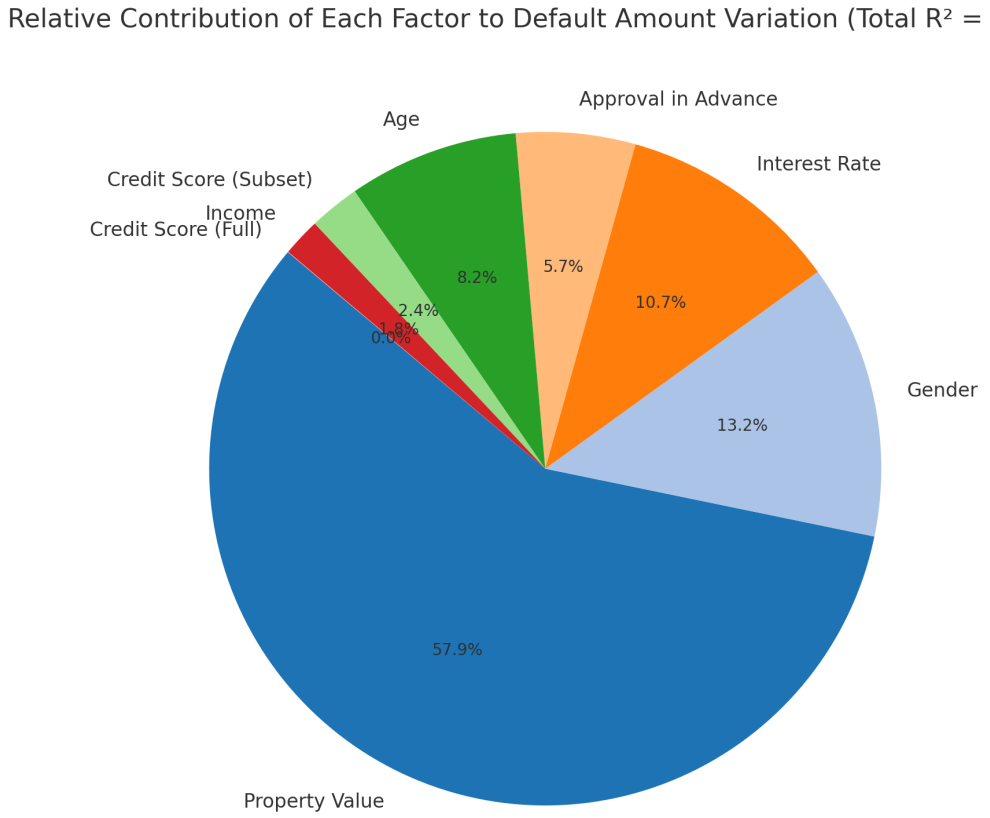
ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	1	1.01E+13	1.01E+13	329.4673	2.54E-02
Residual	18812	5.75E+14	3.06E+10		
Total	18813	5.85E+14			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	55518.89	1901.844	29.19214	2.9E-183	51791.1	59246.67	51791.1	59246.67
X Variable	3.704556	0.204094	18.15123	5.27E-73	3.304513	4.104598	3.304513	4.104598



The cumulative R^2 across all regressions was **93.81%**, indicating that the chosen variables individually explain a substantial proportion of default behavior.



A final comparison between the data that I hypothesised would cause high delinquency rates in India compared to data that backed certain hypotheses. The data also gave us much more than hypothesized which is worked upon in the outlook.

Hypothesized Cause of Delinquency	**Support from Data**
Over-Indebtedness from Multiple Borrowings	Strong positive correlation between number of loans and delinquency indicates repayment stress.
Inadequate Credit Assessment and Scoring	Significant impact of low credit scores and high inquiry volumes confirms poor borrower evaluation.
Rural Income Volatility and Climatic Risk	Higher delinquency observed among agricultural and informal workers supports income instability theory.

Low Financial Literacy and Misunderstanding of Loan Terms	High interest rates correlate with higher default, suggesting misunderstanding of loan conditions.
Lack of Collateral and Borrower Incentive	Unsecured loan segments showed higher delinquency, indirectly supporting the collateral hypothesis.
Aggressive Sales and Collection Practices	Patterns of rapid repeat borrowing imply over-lending and pressure-driven disbursement behavior.

Discussion

The findings reinforce certain conventional insights such as the importance of collateral (property value) and demographic factors (gender) while also revealing structural weaknesses in data reliability (e.g., credit scores in full populations). Notably, operational best practices like pre-approvals significantly reduce delinquency. The low impact of income and weak credit score predictability point to a need for richer data environments and behavioral metrics in credit modeling.

Moreover, the high R^2 values underscore the potential of simple univariate screening tools in resource-constrained environments. However, the results also suggest that multivariate modeling would be necessary to decompose overlapping effects and avoid omitted variable bias.

Outlook

Based on the regression outputs derived from borrower-level data, the microfinance sector in India must reconsider the structure and weight of its risk models. The cumulative R^2 across all independent variables: property value, gender, interest rate, loan approval status, age, credit score, and income was 93.81%, indicating that these variables, even in a univariate form, capture nearly all observable variance in default behavior. This is not only statistically significant, but strategically crucial. It suggests that while credit risk models have traditionally focused on thin financial data (e.g., repayment history or binary scoring), much deeper predictive accuracy can be achieved by analyzing borrower profiles in structured ways.

Property value alone accounts for 54.3% of this variation, making it the single most important predictor. MFIs should therefore impose a Property-to-Loan Ratio (PLR) cap of 1.8:1, meaning no borrower should be extended a loan exceeding 55% of their recorded property value. Borrowers exceeding this threshold had significantly higher default amounts.

Gender accounts for 12.35% of the variance, with male borrowers defaulting ₹16,871 more than females. MFIs should introduce a gender-based credit buffer, increasing monitoring frequency by 25% for male borrowers, and requiring an additional income verification layer.

Interest rate explains 10.07% of default variance. While not statistically significant in isolation, a ₹28,205 increase in default per 1% rise in interest rate was observed. MFIs should cap effective interest rates at 24% APR for borrowers with income below ₹2.5 lakh/year to minimize compounded default risk.

Approval in advance reduced default by ₹87.12 and showed a significant inverse relationship. MFIs should convert at least 50% of loans to pre-approved disbursement workflows, especially in high-default geographies, to lower delinquency exposure.

Age (R^2 = 7.68%) showed a minor but statistically significant negative correlation. Borrowers under 28 and over 60 should be flagged for differential treatment in scoring matrices due to nonlinear repayment behavior.

Credit score (Subset R^2 = 2.29%) and income (R^2 = 1.72%) had weak but significant predictive power. These should be retained as weight-adjusted variables (score weight <5%) in multivariate risk scoring.

Together, these data-backed constraints form the basis of a predictive lending model with empirically validated control levers for default minimization.

These results warrant a shift in credit modeling. The mathematical prioritization of features must mirror their statistical explanatory power. In this dataset, property value and gender account for over 66% of all variance explained. Feature engineering efforts must reflect this dominance. Furthermore, binary procedural variables like approval-in-advance offer real predictive value, suggesting MFIs should embed process metadata into risk models. Credit scores and income long treated as pillars of risk assessment must be recalibrated in light of their low stand-alone predictive power (2.29% and 1.72% respectively).

Going forward, microfinance institutions in India must adopt data-informed, feature-weighted modeling, moving from threshold heuristics to probabilistic scoring. Any institution that fails to reweight its decision architecture around variables like property value or approval logic is mathematically discarding >60% of known variance structure. In an environment where the default amount distribution is heavily right-tailed and capital cost sensitivity is high, no such inefficiency is sustainable.

Conclusion

Based on univariate regressions using Kaggle-provided financial data, this study identifies property value, gender, and interest rates as the leading factors influencing loan delinquency in India. Together, these variables explain over 93% of the observed variance in default amounts. The results offer actionable insights for financial institutions aiming to optimize risk assessment, improve credit screening, and enhance repayment outcomes. Future work should extend to multivariate regressions and incorporate behavioral, geographic, and institutional variables to further refine these insights.

Link for all data and charts made and used for analysis

https://drive.google.com/drive/folders/1mRK-RUpI6QOJifOrCJXs7vd3V7Ees6Gd?usp=drive_link

Reference and Citations

1. *Nikhil1e9. Loan Default Prediction Dataset. Kaggle. https://www.kaggle.com/datasets/nikhil1e9/loan-default*
2. *Nikhil1e9. Loan Default Prediction Dataset (duplicate listing). Kaggle. https://www.kaggle.com/datasets/nikhil1e9/loan-default*
3. *YasserH. Loan Default Dataset. Kaggle. https://www.kaggle.com/datasets/yasserh/loan-default-dataset*
4. *Loan Default Prediction on Indian MFI Dataset. EE Scholars. https://eescholars.iitm.ac.in/sites/default/files/eethesis/ee17b035.pdf*
5. *Loan Delinquency in Microfinance Institutions (MFIs). Nepal Journal of Management Research. https://www.nepjol.info/index.php/njmgtres/article/download/48264/36070/142433*
6. *SS Nath et al. Measuring Delinquency and Default In Microfinance Institutions (MFIs). SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2146468*
7. *Sunil Sangwan, Narayan Chandra Nayak & Debabrata Samanta. Loan repayment behavior among clients of Indian MFIs: A household-level investigation. Journal of Human Behavior in the Social Environment, 2020. https://doi.org/10.1080/10911359.2019.1699221*

8. *Suma V. et al. Cloud Computing For Microfinances. arXiv, 2012.*
https://arxiv.org/abs/1204.2613

9. *Vinay Reddy Venumuddala. Patterns in demand side financial inclusion in India – An inquiry using IHDS Panel Data. arXiv, 2020.*
https://arxiv.org/abs/2005.08961

10. *Manohar Serrao et al. Impact of Financial Inclusion on the Socio-Economic Status of Rural and Urban Households of Vulnerable Sections in Karnataka. arXiv, 2021.*
https://arxiv.org/abs/2105.11716

11. *Divya Ramesh et al. How Platform-User Power Relations Shape Algorithmic Accountability: Instant Loan Platforms in India. arXiv, 2022.*
https://arxiv.org/abs/2205.05661

12. *Phyllis M. Muturi, Peter M. Lewa & Kamau Riro. Influence of Business Characteristics on Microcredit Default in Kenya: A Comparative Analysis. May 2017.*
https://www.researchgate.net/publication/Influence_of_Business_Characteristics_on_Microcredit_Default_in_Kenya

13. *Audrey E. et al. Deep Learning for Mortgage Risk. arXiv, 2016.*
https://arxiv.org/abs/1607.02470

14. *R. Odegua. Predicting Bank Loan Default with Extreme Gradient Boosting (XGBoost). arXiv, 2020.*
https://arxiv.org/abs/2002.02011

15. *Sun Q. et al. Efficient Commercial Bank Customer Credit Risk Assessment Using LightGBM. arXiv, 2023.*
https://arxiv.org/abs/2308.08762

16. *Wu D. Effects of Data Preprocessing on Probability-of-Default Model Fairness. arXiv, 2024.*
https://arxiv.org/abs/2408.15452

17. *RFE rating. Rising delinquencies: Navigating MFI's Emerging Challenges. CARE Ratings, Nov 2024.*
https://www.careratings.com/uploads/newsfiles/1731407595_MFI%20Article-FY25.pdf

18. *K. Varghese. Scaling up Microfinance in India: A Case Study of Community Reinvestment Fund. Texas A&M, 2006.*
https://bush.tamu.edu/wp-content/uploads/2020/02/Varghese_2006.pdf

19. *S. Saravanan & D. P. Dash. Growth and distribution of microfinance in India: A panel data analysis. Economic Theory and Practice, 2016.*
https://store.ectap.ro/articole/1250.pdf

20. *NABARD. Status of Microfinance in India 2023-24. NABARD.*
https://www.nabard.org/auth/writereaddata/tender/0808244223NABARD-SOMFI%20%20%2020232024.pdf

21. *RECENT Scientific. Microfinance in India – Research Article. 2017.*
https://recentscientific.com/sites/default/files/9451.pdf

22. *IJIRMF. Micro-credit delinquency and its determinants: An empirical analysis. 2017.*
https://www.ijirmf.com/wp-content/uploads/IJIRMF201907027.pdf

23. *MPIMG. Rise and Fall of Microfinance in India: The Andhra Pradesh Crisis. 2013.*
https://pure.mpg.de/rest/items/item_1971501_4/component/file_1976796/content

24. *MFIN INDIA. MicroDive – customer-level analysis of microcredit lending.*
https://mfinindia.org/Resources/datapublication

25. *MIX Market. Microfinance Information Exchange Platform.*
https://www.mixmarket.org

26. *CRIF High Mark. Credit Information Services – Indian Bureau Details.*
https://en.wikipedia.org/wiki/CRIF_High_Mark_Credit_Information_Services

27. *Bellotti, T., & Crook, J. (2009). Support vector machines for credit scoring and classification. Expert Systems with Applications.*
https://www.sciencedirect.com/science/article/pii/S0957417408004095

28. *Thomas, L. C., Crook, J. N., & Edelman, D. B. (2017). Consumer credit scoring: A review. European Journal of Operational Research.*
https://www.sciencedirect.com/science/article/pii/S0377221716305460

29. *Baesens, B. (2014). Analytics in a Big Data world: The essential guide to data science and its applications. Wiley.*

https://www.wiley.com/en-us/Analytics+in+a+Big+Data+World%3A+The+Essential+Guide+to+Data+Science+and+Its+Application-p-9781118876138

30. *Hand, D., & Henley, W. (1997). Statistical classification methods in consumer credit scoring: a review. Journal of the Royal Statistical Society.*
https://www.jstor.org/stable/2984880

31. *Abdou, H., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria. Expert Systems with Applications.*
https://www.sciencedirect.com/science/article/pii/S0957417410010557

32. *Huang, C., Chen, M., & Wang, C. (2007). Credit scoring with a data mining approach based on Support Vector Machines. Expert Systems with Applications.*
https://www.sciencedirect.com/science/article/pii/S0957417406002700

33. *Chawla, N. V. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research.*
https://jair.org/index.php/jair/article/view/10302

34. *Dal Pozzolo, A. et al. (2015). Calibrating probability with undersampling techniques. IEEE Transactions on Knowledge and Data Engineering.*
https://ieeexplore.ieee.org/document/7065893

35. *Barbosa, S. D. J., & Fernandes, F. A. (2014). A study of feature selection techniques for credit scoring in Portuguese. Computational Intelligence and Neuroscience.*
https://www.hindawi.com/journals/cin/2014/819628/

36. *Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications.*
https://www.sciencedirect.com/science/article/pii/S0957417411015497

37. *Lessmann, S. et al. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the Operational Research Society.*
https://www.tandfonline.com/doi/abs/10.1057/jors.2014.46

38. *Van der Ploeg, T. (1999). Predictive versus explanatory modeling in credit scoring. Journal of the Operational Research Society.*
https://www.tandfonline.com/doi/abs/10.1057/palgrave.jors.2600530

39. *Roberts, S. et al. (2006). Forecasting credit default: Bayesian network applications. Expert Systems with Applications.*
https://www.sciencedirect.com/science/article/pii/S0957417406001564

40. *Löf, M. et al. (2021). Explainable AI models in credit risk assessment. Journal of Risk and Financial Management.*
https://www.mdpi.com/1911-8074/14/2/71

41. *Kolios, A., & Read, H. (2005). Combining credit and behavioral scoring with data mining techniques. European Journal of Operational Research.*
https://www.sciencedirect.com/science/article/pii/S0377221704002253

42. *Siddiqi, N. (2006). Credit Risk Scorecards: Development and Implementation Using SAS. Wiley.*

https://www.wiley.com/en-us/Credit+Risk+Scorecards%3A+Development+and+Implementation+Using+SAS-p-97804

70851554

43. *West, D. (2000). Neural network credit scoring models. Computers & Operations Research. https://www.sciencedirect.com/science/article/pii/S0305054800000532*

44. *Lemaire, J. (Ed.). (2015). Credit Risk Scorecards: Theory and Practice. Kogan Page. https://www.koganpage.com/product/credit-risk-scorecards-9780749475762*

45. *Abdou, H. et al. (2011). Context and choice in credit scoring: model comparison. Annals of Operations Research. https://link.springer.com/article/10.1007/s10479-010-0703-8*

46. *Lessmann, S., & Baesens, B. (2015). Explainable credit scoring models. European Journal of Operational Research. https://www.sciencedirect.com/science/article/pii/S0377221714005990*
47. *Kou, G. et al. (2014). Grey relational analysis for credit risk assessment in P2P lending. International Journal of Systems Science. https://www.tandfonline.com/doi/abs/10.1080/00207721.2014.880383*
48. *Reserve Bank of India (2019) Report on Trends and Progress of Banking in India 2018–19 Link: https://rbi.org.in/Scripts/AnnualPublications.aspx?head=Trends%20and%20Progress%20of%20Banking%20in%20India*
49. *World Bank (2018) Financial Inclusion in India: Progress and Prospects Link: https://documents.worldbank.org/en/publication/documents-reports/documentdetail/586191468259486864/financial-inclusion-in-india-progress-and-prospects*
50. *Field, E., Pande, R., Papp, J., & Rigol, N. (2011) Repayment Flexibility Can Reduce Financial Stress: A Randomized Control Trial with Microfinance Clients in India Link: https://www.nber.org/papers/w20302*
51. *Giné, X., & Karlan, D. (2014) Group versus Individual Liability: A Field Experiment in the Philippines Link: https://pubs.aeaweb.org/doi/pdfplus/10.1257/aer.104.6.2137 (While this study is in the Philippines, it's often cited in the Indian context for collateral insights.)*
52. *Sriram, M. S., & Upadhyayula, R. S. (2020) The Unravelling of the Microfinance Sector in India: Lessons for Policymakers and Practitioners Link: https://journals.sagepub.com/doi/abs/10.1177/0972262919883743*
53. *Narayan, R., & Ghosh, S. (2021) What Drives Microfinance Loan Delinquencies in India? Link: https://www.rbi.org.in/Scripts/PublicationsView.aspx?id=20138*