

A STATISTICAL FRAMEWORK FOR CONSUMER RISK CLASSIFICATION AND CREDIT UNDERWRITING

Statistics

Anasuri Abhisri	M.Sc. Data Science II Year, Kasturba Degree & PG College for Women, Maredpally, Secunderabad, Telangana, India
Mrs. K Sangeetha	Asst. Professor, Kasturba Degree & PG College for Women, Maredpally, Secunderabad, Telangana, India
Ms. G. Sowmya	Asst. Professor, Department of Statistics, Kasturba Degree & PG College for Women, Maredpally, Secunderabad, Telangana, India

ABSTRACT

Credit risk reflects a borrower's repayment history, obligations and recent credit-seeking behaviour, yet many scoring systems reduce it to a single opaque score. This study developed a CatBoost framework that merges internal bank records with CIBIL bureau data for 51,336 applicants and classifies them into four risk tiers (P1 to P4). A three-stage statistical filter (Chi-Square, VIF and ANOVA) reduced 87 variables to 58, and missing bureau values were kept as an informative signal rather than imputed. On a stratified hold-out set the model reached 80.62% accuracy, with errors confined almost entirely to neighbouring tiers. LIME gave plain-English reason codes, SHAP checked them against overall model behaviour, and a dual-channel pricing engine turned each tier into an interest rate, with a projected net profit of 388.1 million at a 0.10 rejection threshold.

KEYWORDS

Credit Risk Classification; CatBoost; Explainable AI; LIME; SHAP; Informative Missingness; Feature Selection; Dynamic Pricing; RBI Digital Lending

1. INTRODUCTION

Every lending decision rests on one question: how likely is this borrower to repay? Banks long answered it with rigid scorecards built on a bureau score and a few demographic fields, rarely telling applicants how the score was reached (Siddiqi, 2017). Machine-learning models separate good and bad borrowers more accurately (Lessmann et al., 2015), but at the cost of transparency, and a model that cannot explain its decision is hard for a borrower to question or a regulator to audit (Bracke et al., 2019).

Two further problems widen this gap. Many pipelines fill missing bureau fields with an average, yet a blank credit-card utilisation field usually means the applicant holds no card, and a blank delinquency field usually means none is recorded; imputing a mean discards that information. Moreover, most studies stop at classification and never show how predicted risk translates into the interest rate a borrower is offered (Bhatore et al., 2020).

This study addresses both within one framework. Bank and CIBIL records for 51,336 applicants were merged, variables were admitted only after a three-stage statistical filter, and missing values were passed directly to a CatBoost classifier that assigns each applicant to one of four risk tiers. LIME and SHAP explain each decision individually and across the portfolio. On held-out data the model reached 80.62% accuracy, and recent credit behaviour proved a far stronger signal of risk than income.

The predicted tier then feeds a pricing engine that sets the interest rate and approval decision separately for physical banks and digital NBFC lenders, and the pipeline is deployed as a public web application. The focus is therefore the full path from raw data to risk tier, price and an explanation the applicant can understand. The results rest on one bank-bureau dataset and should be read as an internally evaluated research model, not a validated production system.

1.1 Research Objectives

- To develop a classification model that places borrowers into risk tiers to support automated credit approval.
- To examine how the timing and frequency of past defaults relate to current creditworthiness.
- To identify the behavioural variables most strongly associated with default risk.
- To apply Explainable AI so that each credit decision can be accompanied by transparent reasons.

1.2 Hypotheses

H₀: The proposed CatBoost framework does not significantly improve credit risk classification accuracy or explainability over conventional black-box scorecards.

H₁: The proposed CatBoost framework significantly improves credit

risk classification accuracy and explainability over conventional black-box scorecards, enabling safe, transparent underwriting.

2. Related Work

- Ensemble models outperform traditional scorecards (Lessmann et al., 2015), recent behaviour predicts default better than static demographics (Khandani et al., 2010), and CatBoost can use missing-value placeholders directly (Prokhorenkova et al., 2018).
- LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017) give local and global explanations (Bracke et al., 2019), but few of the studies reviewed combine them with auditable feature selection and a pricing layer, which is the gap addressed here.

3. MATERIALS AND METHODS

3.1 Data Source and Description

Bank records and CIBIL bureau records were merged on the applicant identifier (PROSPECTID), giving 51,336 applicants and 87 candidate variables. The target, Approved_Flag, has four ordered tiers: P1 (Excellent), P2 (Good), P3 (Moderate) and P4 (High Risk), shown in Table 2.

3.2 Preprocessing: Informative Missingness

Missing values are stored in the raw data as -99999. The missingness is structured: credit-card and personal-loan utilisation are missing for 92.79% and 86.56% of applicants because most hold no such product, and three delinquency fields for 70.03% because none is recorded (Table 1). The placeholders were therefore converted to true missing values and passed to CatBoost with nan_mode='Min' so the model could learn from absence itself.

Table 1. Top Variables by Missingness Frequency (N = 51,336)

Variable	Count of -99999	% of Dataset
CC_utilization	47,636	92.79%
PL_utilization	44,435	86.56%
time_since_recent_delinquency	35,949	70.03%
max_delinquency_level	35,949	70.03%
time_since_first_delinquency	35,949	70.03%
max_unsec_exposure_inPct	23,178	45.15%
max_deliq_6mts	12,890	25.11%
max_deliq_12mts	10,832	21.10%

3.3 Statistical Feature Selection Pipeline

Three statistical checks decided which of the 87 variables entered the model:

- Stage 1, Chi-Square test: the four categorical fields (MARITAL STATUS, GENDER, last_prod_enq2, first_prod_enq2) were tested against the risk tier; all were significant ($p < 0.05$).
- Stage 2, VIF: numeric variables duplicating other variables (VIF > 6) were removed (O'Brien, 2007).

- Stage 3, ANOVA: numeric variables whose averages differed across the four tiers were retained ($p \leq 0.05$).

3.4 Predictive Model: CatBoost Architecture

CatBoost (Prokhorenkova et al., 2018) was chosen because it handles categorical variables and missing values natively. It was trained on the 58 features with a stratified 80/20 split (41,068 train, 10,268 test) using MultiClass loss, 600 iterations, depth 5, L2 regularisation 5 and nan_mode='Min'.

3.5 Explainable AI Layer: LIME and SHAP

LIME (Ribeiro et al., 2016) converts individual predictions into plain-English reason codes, and SHAP (Lundberg & Lee, 2017) checks that they agree with overall model behaviour.

3.6 Dynamic Pricing and Channel Routing Engine

Predicted tiers feed an RBI-based pricing engine for physical banks and digital NBFCs (Reserve Bank of India, 2022). Base rates run from 9.75% (P1) to 28.00% (P4), plus 1.50% for late-night applications. Physical banks subtract 0.25% and approve only P1 and P2; digital apps add 0.75%, approve P3 only up to ₹50,000 and reject P4.

3.7 System Architecture and Deployment

The pipeline runs as a public Streamlit application with single-applicant underwriting, dual-channel pricing, LIME dashboards and batch scoring.

4. RESULTS & DISCUSSION

4.1 Exploratory Data Findings

Income is strongly right-skewed (skewness 42.20), as are recent enquiries (4.49). Isolation Forest flagged 1,541 of 51,336 records (3.00%) as outliers; they were retained as likely genuine distress rather than errors.

P2 makes up about 62.7% of applicants (Table 2); the imbalance is moderate, so no oversampling was used and 62.7% is the baseline to beat.

Table 2. Target Class Distribution (N = 51,336)

Risk Tier	Risk Category	Share of Applicants
P1 (Class 0)	Excellent	≈ 11.3%
P2 (Class 1)	Good	≈ 62.7%
P3 (Class 2)	Moderate	≈ 14.5%
P4 (Class 3)	High Risk	≈ 11.5%

Kolmogorov–Smirnov tests rejected normality for all variables ($p < 0.0001$; Table 3), supporting a tree-based model.

Table 3. Kolmogorov–Smirnov Normality Tests

Feature	KS Statistic (D)	P-Value (p)	Conclusion ($\alpha = 0.05$)
NETMONTHLYINCOME	0.1909	$p < 0.0001$	Reject H_0 (Non-Normal)
Time_With_Curr_Empr	0.1444	$p < 0.0001$	Reject H_0 (Non-Normal)
Age_Oldest_TL	0.5045	$p < 0.0001$	Reject H_0 (Non-Normal)

4.2 Feature Selection Outcomes

Table 4 shows the weakest variables have IV near zero, while the strongest, Age_Oldest_TL (3.9350) and enq_L3m (3.6129), are behavioural. GENDER passed Chi-Square yet adds almost no predictive value (IV 0.0052).

Table 4. Feature Elimination Via Information Value (IV)

Excluded Feature (Noise)	IV Score	Retained Feature (Top Predictors)	IV Score
num_deliq_6mts	0.0000	Age_Oldest_TL	3.9350
CC_TL	0.0000	enq_L3m	3.6129
GENDER	0.0052	enq_L6m	3.1058
num_sub_6mts	0.0053	time_since_recent_enq	2.8131

4.3 Classification Performance

On 10,268 test records CatBoost reached 80.62% accuracy, well above the 62.72% majority-class baseline (Table 5).

Table 5. Per-tier Classification Performance

Tier	Precision	Recall	F1-score	Support
P1	0.84	0.77	0.80	1,161

P2	0.85	0.94	0.89	6,440
P3	0.48	0.31	0.38	1,491
P4	0.80	0.76	0.78	1,176

Table 6. Confusion Matrix

Actual \ Predicted	P1	P2	P3	P4
P1	897	264	0	0
P2	142	6,033	247	18
P3	34	795	457	205
P4	1	45	239	891

Table 6 shows almost no errors between distant tiers (P1, P4); mistakes stay in neighbouring tiers, mostly at the P2–P3 boundary.

4.4 Explainability Outcomes

Table 7 shows the LIME explanation for a correctly classified P4 applicant.

Table 7. LIME Local Explanation for a P4 (High Risk) Applicant

Factor	Raw Signal	Influence Weight
Recent enquiry volume	enq_L3m > 1.00	+0.4427
Delinquency recency	time_since_recent_delinquency > 10.00	+0.1296
Credit history depth	Age_Oldest_TL ≤ 14.00	+0.1039
Recent standard payments	num_std_6mts ≤ 0.00	+0.0685
Personal loan enquiry share	pct_PL_enq_L6m_of_ever > 0.00	+0.0652

Recent enquiries (“credit hunger”) carry over half of this applicant’s explanation weight (0.4427). SHAP ranks them the top global risk driver, with long credit history and regular recent payments as the main protective factors.

4.5 Deployed Application

The pipeline runs as a public Streamlit application showing live risk tier, dual-channel pricing, LIME explanations and batch scoring (Fig 1). It demonstrates the framework and is not a production lending system.

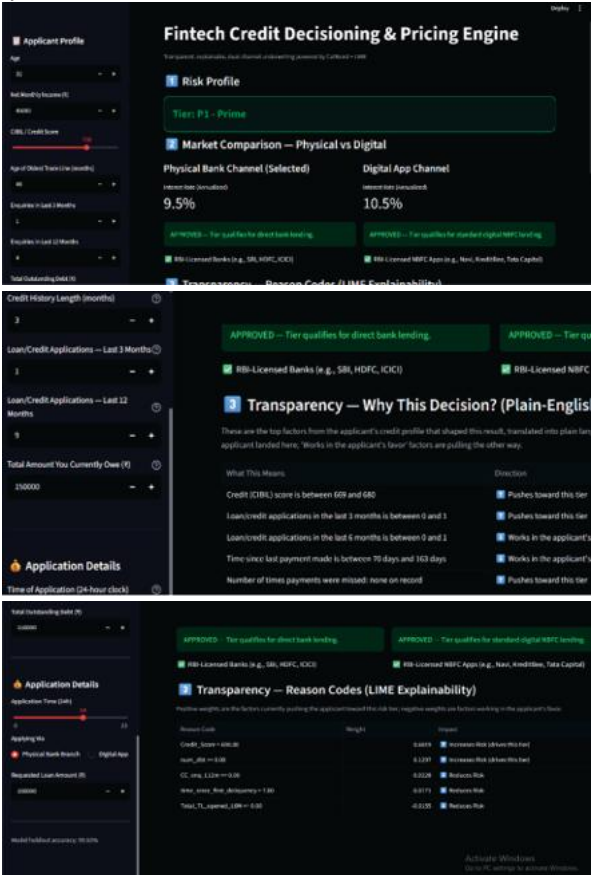


Fig 1. Live Streamlit App Showing Risk Class and Pricing for a Test Applicant.

4.6 Hypothesis Evaluation

H₁: Supported. CatBoost reached 80.62% accuracy against a 62.72% baseline, kept errors within neighbouring tiers, and behavioural variables clearly outperformed demographic ones.

4.7 Behavioral Dominance Over Demographics

Mean recent enquiries rise from 0.39 in P1 to 3.15 in P4 and credit history falls from about 95 to 28 months, while mean monthly income stays between about ₹24,700 and ₹27,300 in all tiers. Lenders relying on demographics and income therefore use comparatively low-information variables, in line with Khandani et al. (2010).

4.8 The P3 Ambiguity Problem

P3 is the main limitation (precision 0.48, recall 0.31): of 1,491 true P3 applicants, 795 were predicted P2 and 205 P4, so P3 acts as a transitional zone that is harder to separate. A specialist model for the P2–P3 boundary is a possible next step.

5. CONCLUSION

This study shows that a credit risk framework can be accurate, statistically justified and transparent within one deployable pipeline. A CatBoost model on 58 selected variables, keeping missing bureau values as information, reached 80.62% accuracy, with almost all errors between neighbouring tiers. Behavioural variables outpredicted demographic ones, delinquency recency added its own weight, and LIME reason codes agreed with SHAP at portfolio level. The RBI-based pricing engine turned tiers into interest rates, with a projected net profit of ₹388.1 million at a 0.10 threshold. The results are internal to one dataset, P3 remains weakest, and the pricing rules are untested against actual repayment, so independent validation is needed.

REFERENCES

1. Bhatore, S., Mohan, L., & Reddy, Y. R. (2020). Machine learning techniques for credit risk evaluation: A systematic literature review. *Journal of Banking and Financial Technology*, 4(1), 111–138.
2. Bracke, P., Datta, A., Jung, C., & Sen, S. (2019). Machine learning explainability in finance: An application to default risk analysis (Staff Working Paper No. 816). Bank of England.
3. Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11), 2767–2787.
4. Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1), 124–136.
5. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
6. O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, 41(5), 673–690.