Rath and Date *EPJ Data Science* (2026) 15:44
https://doi.org/10.1140/epjds/s13688-026-00655-0

EPJ Data Science

RESEARCH Open Access

A row-type specific hybrid framework for credit risk analysis: loan portfolio based feature selection and unsupervised Bayesian network dependency exploration

Minati Rath[1*] and Hema Date[1]

Handling Editor: David Garcia

*Correspondence:
minati.rath.2019@iimmumbai.ac.in
[1]Department of Analytics and Decision Science, Indian Institute of Management (IIM), Mumbai, India

Abstract

Credit risk assessment is a critical function in financial analytics, requiring models that can adapt to diverse borrower profiles while providing clear and interpretable insights. Although a range of data driven techniques have been applied in this domain, many struggle to handle the inherent heterogeneity of financial data across different loan categories such as personal and agricultural loans. This paper introduces `Credit Risk Analysis with Bayesian Networks (CRAB-Net)`, a row type specific hybrid framework for credit risk modeling. The approach first segments the data by loan type and balances the distribution of risk categories to ensure fair representation. It then identifies the most outcome relevant attributes through targeted feature selection, focusing on variables most associated with credit risk differentiation. On this refined set of features, unsupervised Bayesian network learning is applied to uncover conditional dependencies among financial variables without relying on default outcome labels. This design combines supervised relevance filtering with unsupervised dependency discovery, reducing noise and avoiding misleading patterns from analyzing all features indiscriminately. The framework revealed that in personal loans, installment-related variables such as installment frequency, overdue status, and repayment structure emerged as central nodes, indicating their dominant role in defining repayment behavior and delinquency risk. In contrast, for agricultural loans, the network structure was shaped primarily by provisioning norms, landholding details, and exposure-related attributes such as sanctioned amount and collateral type, suggesting that borrower risk in this segment is more closely linked to regulatory classification and collateral strength. Experiments on real world banking data show that `CRAB-Net` provides interpretable dependency graphs, supports fair segment level analysis, and enhances transparency for audit and supervisory compliance under Basel norms by offering clear, data driven evidence of the risk factors shaping borrower outcomes.

Keywords: Credit risk analysis; Personal loan prediction; Agriculture loan prediction; Bayesian network

1 Introduction

Credit risk assessment is a foundational task in financial decision making, directly influencing loan approvals, pricing strategies, and portfolio management [1]. As financial institutions increasingly adopt data driven systems for evaluating risk, the need for robust, interpretable, and adaptable modeling frameworks has become critical [2–4]. Traditional credit scoring approaches and supervised machine learning models, while widely used, often rely on the assumption of homogeneity in borrower and loan characteristics [5]. Prior studies have highlighted this limitation in conventional credit risk modelling, where pooled borrower data are treated as structurally uniform, potentially masking portfolio-specific drivers of default [6–8]. These approaches optimise predictive accuracy across combined datasets but seldom account for how repayment dynamics, exposure risk, and collateral behaviour differ across loan categories or borrower types. In practice, financial datasets are highly heterogeneous, particularly across distinct loan categories such as personal and agricultural loans, which presents substantial challenges to generalization and interpretability.

Most existing models prioritize predictive accuracy based on labeled outcomes, but they frequently fail to uncover deeper structural dependencies among financial variables. On the other hand, purely unsupervised methods that seek to learn patterns from all available features risk incorporating attributes irrelevant to credit risk, leading to misleading or noisy dependency structures. These issues are especially consequential in regulated environments, where frameworks such as Basel II and III require institutions to provide clear, data driven justifications for credit decisions.

To address these challenges, this study proposes a row type specific hybrid framework that combines supervised and unsupervised learning to capture both outcome relevance and underlying conditional dependencies. The methodology first segments the dataset by loan type to preserve the unique characteristics of each portfolio. We refer to this as "row type specific modeling", where each data row is assigned to a loan category such as personal or agricultural and modeled within its segment. This approach allows the framework to reflect the distinct credit structures and behavioral patterns associated with each loan portfolio. Within each segment, `Synthetic Minority Over-sampling Technique (SMOTE)` is applied to balance class distributions, followed by `XGBoost Classifier (XGBoost)` based supervised feature selection to identify the most predictive variables with respect to credit risk outcomes. By restricting the subsequent analysis to these outcome relevant features, the framework minimizes the inclusion of spurious relationships that can arise in fully unsupervised analyses. On this refined feature space, we then apply unsupervised Bayesian network structure learning to explore conditional dependencies without using labels during this stage. This three stage process results in directed graphical models that reveal how key financial attributes interact, supporting detailed analysis through Markov blanket extraction around critical nodes.

The proposed framework, termed `CRAB-Net`, enables scalable and interpretable analysis tailored to heterogeneous financial portfolios. It provides insights into dependency structures unique to each loan type, facilitates differentiated credit strategies, and strengthens compliance with prudential standards by offering explicit graphical evidence of risk drivers. By integrating supervised and unsupervised components in a row type specific pipeline, `CRAB-Net` advances the development of transparent, data driven credit risk assessment practices aligned with regulatory expectations.

2 Research contribution

This study offers several important contributions to the field of creditrisk assessment:

Row Type Specific Hybrid Modeling: We introduce an adaptive framework that integrates row type specific data segmentation with a two stage learning process. It first applies supervised feature selection using `SMOTE` and `XGBoost` to identify outcome relevant variables within each loan category, followed by unsupervised Bayesian network structure learning to uncover deeper conditional dependencies.

Probabilistic Dependency Structures: By employing Bayesian networks on the feature sets refined through supervised selection, we reveal conditional relationships among financial variables that are highly relevant to credit outcomes. This generates transparent graphical models that facilitate detailed analysis through Markov blanket extraction and enhance interpretability of localized risk drivers.

Reduction of Spurious Patterns: The combination of supervised and unsupervised techniques ensures that dependency exploration is focused on the most risk relevant attributes, minimizing the inclusion of misleading relationships that often arise in purely unsupervised analyses.

Application to a Real World Banking Dataset: We validate the `CRAB-Net` framework on a comprehensive banking dataset containing both personal and agricultural loans. This demonstrates its capability to differentiate risk structures across heterogeneous loan types in a manner not addressed by traditional supervised models alone.

Scalable and Transparent Decision Support: The resulting system provides a scalable and interpretable approach for credit risk management. It supports data driven decision making and aligns with regulatory requirements for model transparency under standards such as Basel II and III, strengthening internal auditability and compliance.

3 Related work and literature review

Credit risk modeling has long relied on supervised machine learning algorithms such as logistic regression, decision trees, and ensemble methods [9, 10]. These models typically depend on labeled outcomes, such as loan defaults, and often assume homogeneity in borrower behavior across loan types [11, 12]. While these approaches are effective in static and well labeled environments, their performance often declines in real world settings where data scarcity and heterogeneity are common [13].

With the increasing availability of unlabeled data, unsupervised techniques such as clustering, including K-means and DBSCAN, principal component analysis, and autoencoders have been applied to detect anomalous behavior and latent structures in financial data [14–16]. Although clustering and dimensionality reduction methods can reveal hidden patterns, they do not explicitly represent conditional dependencies or provide interpretable graphical structures that support causal reasoning. Graphical lasso and Markov random fields have been used in financial anomaly detection, but these primarily capture undirected associations, which limits their use for directional interpretations required in policy and credit strategy development.

Bayesian probabilistic learning has received significant attention in financial risk modeling because it can incorporate uncertainty into predictive frameworks [17, 18]. Bayesian networks offer a graphical modeling approach that captures conditional dependencies among variables and supports probabilistic reasoning in decision making [17, 19, 20] [21, 22]. Previous work has applied Bayesian networks to credit scoring, but most studies rely on labeled outcomes or use fixed structures, which limits adaptability to heterogeneous or unlabeled environments. Recent studies have explored structure learning to uncover latent relationships, which is useful in unlabeled or semi supervised contexts. Pham et al. [23] proposed an unsupervised training approach for Bayesian network classifiers applied to data clustering. Their work evaluated models such as Chow Liu trees and Tree Augmented Naive Bayes, optimized through maximum likelihood estimation within the expectation and maximization framework. Their work focused on clustering to improve structure learning but did not integrate segment specific probabilistic modeling aligned with loan types. Zhang et al. [24] introduced a two step clustering strategy that first divides nodes into clusters and then learns structures within and across these clusters. Gal et al. [25] proposed dropout as a Bayesian approximation to model uncertainty in deep learning, improving risk estimation accuracy. Blundell et al. [26] explored weight uncertainty in neural networks using variational inference, which improved interpretability in financial applications. Kypriotakis et al. [27] applied Dirichlet process mixture models for financial risk assessment, showing the advantages of Bayesian nonparametrics in managing heterogeneous data.

Recent literature underscores the regulatory need for transparent and well documented credit risk models under frameworks such as Basel II and III. These frameworks require banks to maintain adequate capital buffers based on documented risk assessments [28, 29]. Studies such as [30] and [31] explain how supervisors expect financial institutions to justify credit decisions using structured, data driven analyses, particularly during stress testing. Research in explainable AI has responded to these requirements by developing methods that improve model transparency. Ribeiro et al. studied local explanation methods for complex classifiers [32], while Rudin argued that inherently interpretable models should be used in financial decision processes [33]. Bussmann et al. and Misheva et al. provided explainable AI frameworks specifically for credit scoring and risk management [34, 35]. Network based methods have also been applied to assess interconnected exposures and systemic risk, as demonstrated by Hurd in studies of credit contagion in banking networks [36]. These contributions show the importance of generating explicit probabilistic or network structures that supervisors and auditors can review.

Row-type dependent analysis provides an approach that accounts for variations in attribute behavior across different loan categories. Related frameworks such as context aware learning and multi task learning also adapt models to specific data patterns [37–39]. Unlike traditional approaches that assume homogeneous dependencies, row-type segmentation tailors structures to each loan type and improves both accuracy and interpretability. Previous work has highlighted the importance of row-type models in banking [40], while studies on quantum powered row-type frameworks have demonstrated their ability to capture complex dependencies [41].

Despite these advances, no previous work combines supervised feature selection with unsupervised Bayesian network learning in a row-type-specific pipeline for credit risk analysis. While the CRISP-DM framework [42] conceptually supports iterative feedback loops between its stages, many domain-specific implementations of credit risk modeling still adopt a largely sequential and globally homogeneous approach that applies the same modeling logic across loan segments. Methods such as those proposed by Pham et al. [23] and Zhang et al. [24] perform clustering prior to Bayesian network learning but do not leverage supervised outcomes to refine features beforehand, nor do they explicitly segment by loan type within a regulatory context. By first applying `SMOTE` and `XGBoost` for label-guided feature selection and then conducting unsupervised `Bayesian Network (BN)` learning, our framework ensures that the explored dependency structures are both outcome-relevant and unbiased by label information during the graphical analysis stage.

This hybrid approach enables a more targeted exploration of latent conditional dependencies, reduces the risk of capturing irrelevant patterns, and directly supports the need for clear, auditable rationales under Basel II and III. It systematically uncovers nuanced dependency patterns across personal and agricultural loans, offering a novel contribution to the literature on interpretable, heterogeneous credit risk modeling.

4 Methodology: `CRAB-Net` framework

This study is structured to uncover the financial drivers and interdependencies that shape credit risk across a diverse lending portfolio. Unlike traditional supervised machine learning approaches that emphasize prediction, our objective is to achieve a deeper understanding of how borrower, loan, and account characteristics jointly influence repayment outcomes under different credit products. This perspective is particularly important in banking environments that require transparency, auditability, and alignment with prudential regulations such as Basel II and III.

We note that frameworks such as CRISP-DM [42] encourage iterative feedback loops between business understanding, data understanding, data preparation, modeling, and evaluation. While these loops allow for accommodating structural heterogeneity, in many practical applications, modeling is often applied uniformly across loan segments. Our approach, Row-Type Dependent Predictive Analysis (RTDPA), explicitly incorporates heterogeneity-aware modeling steps by tailoring preprocessing, feature selection, and learning to each loan segment, thereby addressing segment-specific risks more effectively. This positions the CRAB framework as a heterogeneity-sensitive implementation, even if the current illustration shows a primarily sequential flow.

We implement this through `CRAB-Net`, which stands for Credit Risk Analysis with Bayesian Networks. The methodology as shown in Fig. 1 begins with data preprocessing for each loan type. The comprehensive heterogeneous dataset is segmented into separate groups for personal and agricultural loans along with their relevant features. For instance, attributes such as `WETLAND, DRYLAND`, which have meaning primarily in the context of agricultural lending, are preserved only for that segment. This segmentation prevents misleading results that could occur if fundamentally different loan types were forced into a single model structure.

For each loan segment, we prepare the data by handling missing values, treating outliers, and normalizing financial fields to support fair analysis. Under risk class balancing, we apply `SMOTE` to adjust for the typical imbalance [43], where standard loans greatly outnumber substandard, doubtful, or loss cases. This ensures that patterns tied to higher risk categories are properly represented in the subsequent analysis.

In the feature reduction step, we apply Principal Component Analysis `(Principal Component Analysis (PCA))` separately on each loan segment. This technique reduces the dimensionality of the data, simplifying the analysis by focusing on the primary

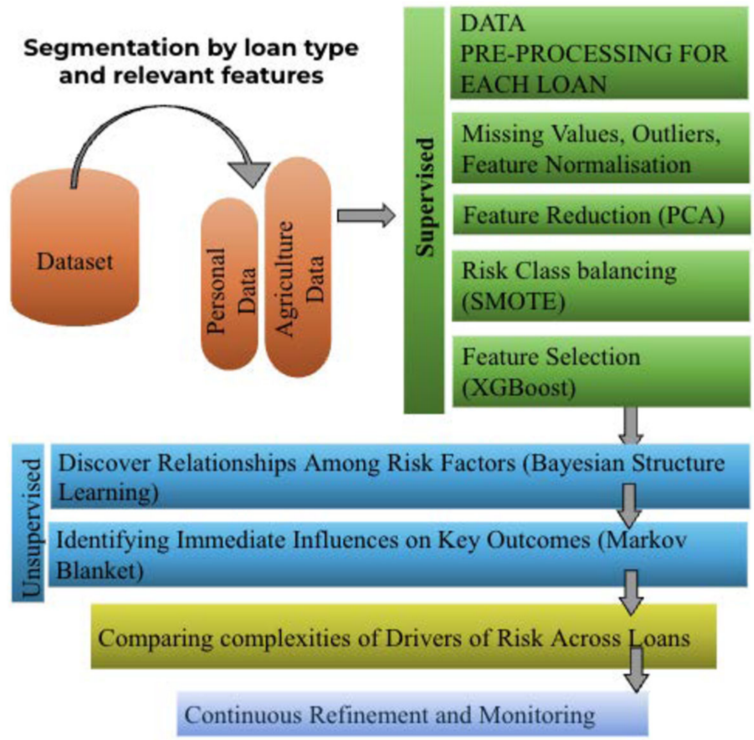


Figure 1 CRAB-Net Methodology Framework

financial patterns within each loan type. Because personal and agricultural loans exhibit distinct characteristics and variability, the PCA process results in different sets and numbers of key attributes for each segment. This ensures that the subsequent analysis is tailored to the dominant financial drivers specific to each loan type, rather than imposing a common structure that could obscure meaningful differences.

Importantly, rather than applying XGBoost directly on the PCA components, the retained components are mapped back to the original feature space using their exact factor loadings. This mapping identifies the original variables that contribute most strongly to the variance directions captured by PCA, allowing XGBoost to operate on genuine variables while retaining the denoising and decorrelation benefits of PCA. By using these loadings, the Bayesian Network is built on genuine financial variables rather than abstract linear combinations of features. This preserves interpretability, allowing the framework to provide transparent insights into the key drivers of credit risk. This preserves interpretability while also reducing multicollinearity and ensuring that feature importance scores are stable and reliable.

Following this mapping, XGBoost is applied to the selected original features to rank them by importance. This ensures that only variables most relevant to differentiating credit risk categories are retained for unsupervised Bayesian network learning. Features with zero or negligible importance are excluded, preventing noise from globally irrelevant variables from affecting the network structure. This approach maintains both interpretability and robustness of the selected features.

This hybrid approach is preferable to alternative methods. Using PCA-only components directly would result in Bayesian Network nodes representing linear combinations of variables, undermining interpretability. Using XGBoost-only alone on all features would be sensitive to noise and multicollinearity, potentially producing unstable or redundant feature sets. By combining PCA and factor loading for denoising and decorrelation with XGBoost for feature ranking, the methodology achieves a balance of interpretability, stability, and relevance for the subsequent unsupervised analysis.

Next, we perform unsupervised Bayesian network learning on each loan type to discover relationships among risk factors. The Bayesian network structure for each loan portfolio was derived using a score-based approach implemented with the Hill Climb Search algorithm. The Bayesian Information Criterion (BIC) was used as the scoring function to balance model complexity with data fit. No edge directionality was imposed manually; the algorithm inferred directed acyclic graphs based purely on statistical dependencies within the selected feature subset for each portfolio. No expert constraints, temporal ordering, or domain rules (such as IRAC compliance) were injected during structure learning to preserve the unsupervised nature of dependency discovery. This design choice allowed the framework to reveal latent patterns and structural asymmetries without presupposing domain-specific relationships, though the resulting networks were later interpreted in light of regulatory logic and business relevance. This step creates probabilistic graphs that map how key variables, such as sanctioned loan amounts, installment counts, interest accumulation, provisioning totals, and customer outstandings, influence each other. To further understand these interactions, we extract the Markov blanket around important financial measures like `CUSTOTOUT`, identifying the immediate set of features most directly connected to a customer's outstanding exposure. The complete Markov blankets for all principal financial variables are documented in the Appendix A.1, providing a comprehensive reference to support transparency, managerial review, and audit processes.

We then compare these dependency structures across loan types. Personal loans show more interconnected patterns among financial attributes, while agricultural loans involve simpler relationships. This suggests the need for more nuanced monitoring and tailored strategies for personal lending, whereas agricultural portfolios may be managed with more focused risk checks.

Finally, under continuous refinement and monitoring, `CRAB-Net` supports periodic updates to these dependency structures as new data becomes available. This allows the institution to maintain an accurate understanding of evolving credit risk patterns and ensures that credit strategies, provisioning, and capital planning stay aligned with the actual dynamics in the loan portfolio.

The hybrid approach followed in this study is essential for practical credit risk analysis. A purely unsupervised exploration over all available attributes would likely capture many variables unrelated to default risk, resulting in noisy or misleading dependency structures. On the other hand, using only supervised prediction could miss the deeper ways in which financial features influence each other beyond direct default outcomes. By first balancing risk categories and identifying the most critical variables with supervised methods, and then applying unsupervised Bayesian network learning to explore how these variables interact, the framework achieves both targeted relevance and an understanding of conditional relationships. This combination supports transparent, data-driven decisions and helps meet regulatory expectations for clear documentation of the factors that drive credit risk.

CRAB-Net pipeline Credit Risk Analysis with Bayesian Networks

Notations:

D: Input dataset with N instances and M attributes.

R: Set of distinct row types in D (for example personal, agricultural loans).

D_r: Subset of D containing type specific rows and relevent features of type $r \in R$.

X_r: Feature matrix for row type r.

1. Row-Type specific Data Segmentation and Preprocessing
   Segment the dataset D into subsets D_r based on row types r.
   Perform row-type specific preprocessing, handling missing values, outliers, and feature normalization.
2. Risk Class Balancing and Feature Selection
   Apply `SMOTE` to each D_r to balance the credit risk categories.
   Use `PCA` on the balanced data to reduce dimensionality while retaining most of the variance.
   Employ `XGBoost` on the `PCA`-transformed data to identify and retain the most important features linked to credit risk.
3. Unsupervised Structure Learning using Bayesian Networks
   For each row type r:
   Construct Bayesian Network G_r using unsupervised structure learning algorithms on X_r.
   Optimize network structure to capture conditional dependencies among features.
4. Dependency Graph and Heatmap Generation
   Visualize G_r as directed graphs and derive dependency matrices for heatmaps.
   Compare interaction strengths across row types.
5. Markov Blanket and Conditional Influence Analysis
   Extract Markov blankets for critical variables such as `CUSTOTOUT`.
   Interpret localized conditional influences driving risk.
6. Cross Type Structural Analysis
   Compare network structures across personal and agricultural loans to identify differing dependency patterns.
7. Periodic Adaptation and Operational Deployment
   Update Bayesian structures as new data becomes available.
   Enable ongoing, interpretable monitoring of risk profiles tailored by row type.

To maintain clarity in the main exposition, detailed Markov blanket structures and variable dependency heatmaps are presented in Appendix A.1 and 7. These supplementary materials provide technical support for the interpretations discussed in the results sections.

5 Experimental results

As this study does not involve supervised prediction, we do not benchmark against traditional classifiers such as logistic regression or scorecard models. The focus is structural discovery, not predictive performance.

5.1 Data

Our dataset comprises over 25,000 samples sourced from a commercial bank in India, containing 81 attributes detailed in Table 12. Loans are grouped into two primary cat-

Table 1 Loan Classification without RTDPA

IRAC	Count	IRAC	Count
Standard	21,894	Sub standard	420
Doubtful	2706	Loss	213

Table 2 Loan Classification with RTDPA

Loan Type	IRAC	Count
Agriculture Loan	Standard	17,496
	Sub standard	294
	Doubtful	2577
	Loss	210
Personal Loan	Standard	4398
	Sub standard	126
	Doubtful	129
	Loss	3

Note. IRAC norms as per RBI: *Standard* — overdue <90 days; *Substandard* — 90 days–12 months; *Doubtful* — >12 months; *Loss* — identified as non-recoverable

egories: Agriculture loans and Personal loans. Each loan is classified under one of four Income Recognition and Asset Classification (IRAC) categories: standard, substandard, doubtful, or loss, as summarized in Table 1 and 2. Notably, in the personal loan segment, there are only three samples categorized as "loss," prompting us to consolidate them into the "doubtful" category for meaningful analysis.

The dataset captures each loan as a single record at the time of disbursement. The data covers loans disbursed over a multi-year period, representing a snapshot of the bank's credit portfolio during this timeframe. We segregated the dataset by loan type to allow tailored processing and modeling for Agriculture and Personal loans.

5.2 Data PreProcessing

We performed data preprocessing separately for each loan type to preserve row-type specific characteristics. For personal loans, features such as `DRYLAND` and `WETLAND` were not applicable and were therefore removed.

We then analyzed missing values and dropped features with more than 70% missing data, as listed in Table 3.

Table 3 shows that several variables have very high levels of missing data in both datasets, often exceeding 90%. Variables such as `OPINIONDT`, `DOCREVDT`, `PRISECCD2`, `RENEWALDT`, and `INSEXPDT` exhibit consistently high missing proportions across personal and agricultural loans.

Some variables are loan-type specific due to differences in operational and regulatory requirements for each loan category. For instance, `DIRFINFLG` appears only in personal loans, while `UNIFUNFLG` is recorded only for agricultural loans where certain funding schemes may apply. Additionally, variables such as `TFRDT`, `RECALLDT`, and `WOSACD` correspond to exceptional administrative actions (e.g., loan transfer or recall), which occur only for a small subset of accounts. Due to their extremely high sparsity, these variables were removed from further analysis.

Significance of row-type specific handling Our approach, based on row-type segmentation, ensured that preprocessing steps were tailored to the unique characteristics of per-

Table 3 Variables with More Than 70% Missing Values for Personal and Agriculture Loans

Variable	Description	Personal Loan		Agriculture Loan	
		Total Missing	% Missing	Total Missing	% Missing
OPINIONDT	Opinion Date	4436	95.3	19,743	95.9
DIRFINFLG	Director Financial Flag	4656	100.0	-	-
SANAUTCD	Sanction Authority Code	3838	82.4	18,067	87.8
DOCREVDT	Document Received Date	4229	90.8	19,918	96.8
PRISECCD2	Primary Security Code 2	4406	94.6	19,668	95.6
RENEWALDT	Renewal Date	4656	100.0	19,956	97.0
INSEXPDT	Insurance Expiry Date	4439	95.3	20,502	99.6
TFRDT	Transfer Date	4216	90.5	18,090	87.9
REASONCD	Reason Code	4181	89.8	18,081	87.9
RECALLDT	Recall Date	4216	90.5	18,096	87.9
WOSACD	Work Order / Service Code	4568	98.1	19,579	95.1
UNIFUNFLG	Uniform Fund Flag	-	-	15,771	76.6

sonal and agricultural loans. After removing highly sparse variables, the remaining features were processed separately for each loan type to preserve meaningful structural differences between the datasets.

For personal loans, missing values were primarily associated with borrower-specific or documentation-related attributes. In contrast, agricultural loans showed missingness patterns linked to sector-specific operational processes. Handling these datasets separately allowed us to maintain relevant signals while avoiding distortions that could arise from applying a uniform preprocessing strategy.

This row-type aware preprocessing improved overall data quality and supported more reliable downstream analysis.

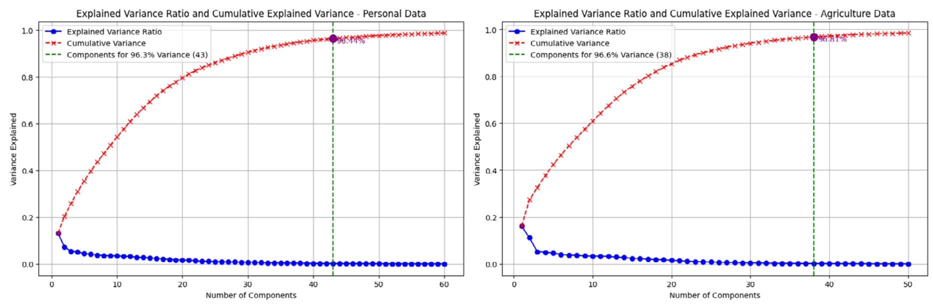
The original dataset exhibits significant class imbalance, especially with a large number of standard loans compared to substandard, doubtful, and loss cases. To address this, we then applied `SMOTE` to synthetically balance the minority classes. This balancing was performed separately for agriculture and personal loans, ensuring that each segment preserves its row-type specific characteristics.

5.3 Feature reduction with PCA

After completing risk class balancing for each loan type, we addressed the high dimensional nature of the datasets through `PCA`. By transforming the original correlated variables into a new set of orthogonal principal components, `PCA` made it possible to capture the key variance directions in a reduced feature space, simplifying further analysis without significant information loss.

To decide how many components to retain, we reviewed the scree plots for both loan types, shown in Fig. 2. The plots showed a clear point after which adding more factors provided very little additional insight into the overall variation in the data. Based on this, we retained 38 principal components for agricultural loans, capturing 96.6 % of the total variance, and 43 components for personal loans, which accounted for approximately 96.3% of the variance.

These results highlight that while agricultural loan data structure is dominated by fewer strong factors, personal loans distribute variance across a slightly broader range of dimensions. This suggests more diffuse patterns of influence in personal lending, whereas agricultural loans rely more heavily on a smaller set of key attributes. By reducing the data



(a) PCA applied to personal loan applicant data (b) PCA applied to agriculture loan applicant data

Figure 2 Dimensionality reduction results using PCA. The left plot shows PCA transformed feature distribution for personal loan applicants, while the right plot shows the same for agriculture loan applicants. The transformation helps visualize latent structure in high dimensional input data

to these principal components, we maintained nearly all of the original variability, laying a compact yet robust foundation for downstream analysis.

5.4 Feature selection using `XGBoost`

Following the exploratory PCA checks, we applied XGBoost for label guided feature selection. This process enabled us to identify and retain the features most strongly associated with the credit risk categories (standard, substandard, doubtful, loss) within each loan type. The model was not tuned for optimal prediction, and default hyperparameters were used to avoid overfitting or segment specific bias. The purpose of XGBoost here was not supervised prediction but label informed relevance filtering prior to unsupervised Bayesian network modeling. By narrowing the feature set based on outcome relevance, we ensured that the subsequent unsupervised Bayesian network structure learning would focus on the most informative attributes. Importantly, after this stage, all labels were discarded, and the Bayesian networks were learned purely from the selected features' internal statistical relationships.

PCA reduced the dimensionality of X_r from $M > 80$ to 43 (personal) and 38 (agricultural), lowering noise and complexity. XGBoost then selected < 20 outcome relevant features from X_r, enabling interpretable and tractable Bayesian network learning on G_r.

5.5 Relationships among risk factors (Bayesian network learning and structural insights)

Using the balanced and feature-selected datasets, we performed unsupervised Bayesian network structure learning separately for agriculture and personal loans. This produced directed acyclic graphs that reveal how attributes jointly influence credit risk within each segment. In the network visualizations, edges are color-coded to indicate directionality: the blue portion of an edge corresponds to the originating factor, while the pink portion corresponds to the target factor. This coloring scheme helps readers visually trace the flow of conditional dependencies between financial risk attributes. The dependencies among variables are not all of equal magnitude; each conditional relationship varies according to the data. In the figures, edge thickness is not used, and the focus is on the presence and direction of relationships rather than their quantitative strength. In the Bayesian network paths, the start variable is selected based on the network structure as a factor that

Table 4 Variables Used as Nodes in the Personal and Agricultural Bayesian Networks

Variable	Description	Personal	Agriculture
BORWORAMT	Borrower Outstanding Amount	✓	✓
CUSTOTOUT	Customer Total Outstanding Amount	✓	✓
DISTTCD	District Code	✓	✓
DPAMT	Down Payment / Drawing Power Amount	✓	✓
DRYLAND	Dryland Landholding Indicator	–	✓
INSAMT	Insurance Amount	–	✓
INSTALAMT	Installment Amount	✓	✓
NFMRGAMT	Non-Farm Margin Amount	✓	–
NUMOFINST	Number of Installments	✓	✓
ORIGINAMT	Original Sanctioned Amount	✓	✓
PERIOD	Loan Period / Tenure	✓	✓
PRISECAMT	Primary Security Amount	✓	✓
RECAMT	Recovery Amount	✓	✓
SBWCLMTAMT	Sanctioned Borrower Working Capital Limit Amount	✓	✓
SBTLLMTAMT	Subtotal Loan Limit Amount	✓	✓
SPLCSAMT	Special Case Subsidy Amount	✓	✓
SUBSIDYAMT	Subsidy Amount	–	✓
SUTFILAMT	Suit Filed / Subsidy Utilization Amount	–	✓
TPGAMT	Total Proposed General Amount	✓	–
WETLAND	Wetland Landholding Indicator	–	✓
YTDCSUMAMT	Year-to-Date Credit Summation Amount	✓	✓
YTDINTAMT	Year-to-Date Interest Amount	✓	✓

influences one or more downstream variables. A path follows directed edges through conditional dependencies from the start variable to subsequent variables. While some paths end at variables with no further dependencies, others continue through intermediate nodes according to the network connections. This construction highlights how upstream risk factors propagate their influence through the network. To improve clarity for the reader, each path is presented in a table showing the start variable, the full sequence of conditional dependencies, and the managerial insight. Edges in the Bayesian networks represent conditional dependencies between financial risk factors. An edge originating from a variable (shown as the blue portion) and pointing to another variable (pink portion) indicates that the target variable is conditionally dependent on the source variable, given the rest of the network. This directional relationship allows readers to trace how changes in one factor may influence others within the credit portfolio. Table 4 presents the variables that emerge as nodes in the Bayesian networks learned from the data for personal and agricultural loans, along with their descriptions and their presence in each network.

5.5.1 Personal loan Bayesian network

The Bayesian networks for personal and agricultural loans represent credit risk relationships among financial variables, where nodes correspond to financial risk factors and edges denote probabilistic dependencies. A comparative analysis of these networks provides insights into their complexity, key influencers, and financial risk patterns.

The personal loan network, as shown in Fig. 3 comprises 18 nodes and 71 edges, indicating a highly interconnected structure with multiple dependencies among financial attributes. One of the most influential nodes in this network is SBWCLMTAMT (Sanctioned Borrower Working Capital Limit Amount), which exhibits direct or indirect condi-

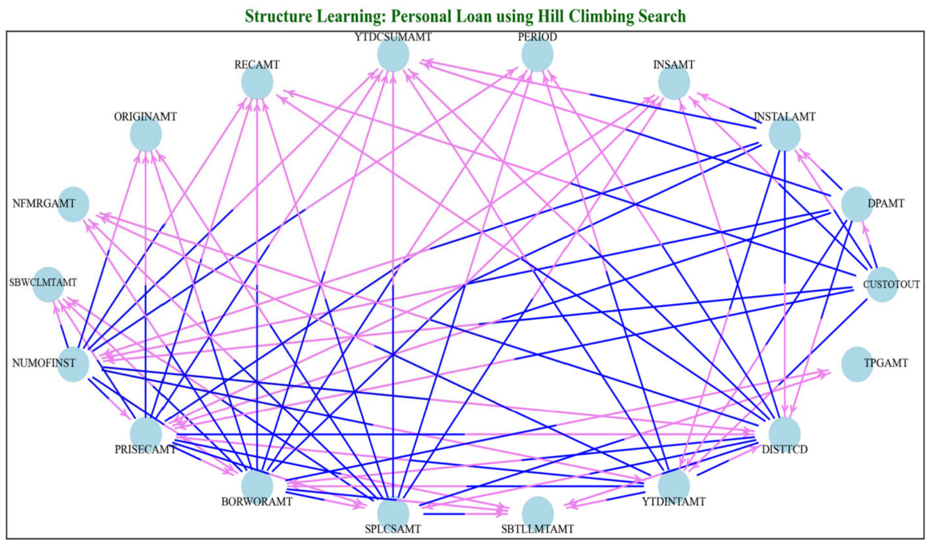


Figure 3 presents the Bayesian network structure automatically learned for the personal loan portfolio. As the graph is generated algorithmically, manual modification of its layout is not feasible. Each node represents a selected financial attribute, and directed edges indicate conditional dependencies inferred using BIC based scoring. To aid interpretation, we focus the discussion on a few key relational patterns—particularly those involving highly connected nodes such as SBWCLMTAMT (Sanctioned Borrower Working Capital Limit Amount), NUMOFINST (Number of Installments), and YTDINTAMT (Year-to-Date Interest Amount). These nodes collectively define the dominant dependency flow (SBWCLMTAMT → NUMOFINST → INSTALAMT → DPAMT), which captures the essential structure of repayment relationships within the personal loan segment. See Table 12 for full variable label mapping

tional dependencies with several key attributes, including NUMOFINST (Number of Installments), PRISECAMT (Primary Security Amount), and BORWORAMT (Borrower's Outstanding Amount). These connections are reflected in the dependency chain SBWCLMTAMT → SPLCSAMT → BORWORAMT → DISTTCD → PRISECAMT → YTDINTAMT → NUMOFINST → INSTALAMT → DPAMT (Table 5) and confirmed through the Markov Blanket Membership Matrix A.1, where these variables co-occur within the local dependency neighborhood of SBWCLMTAMT. This clarifies that the sanction limit node influences repayment and exposure variables through a multi-step propagation of conditional dependencies rather than purely direct edges, emphasizing its structural dominance in the personal loan network. The high connectivity of these variables suggests that sanctioned loan amounts significantly impact repayment structures and financial commitments.

The network also highlights the importance of `NUMOFINST`, which directly influences `CUSTOTOUT` (Customer Outstanding Amount), `DPAMT` (Down Payment Amount), and `INSTALAMT` (Installment Amount). This interdependency suggests that the number of installments plays a crucial role in determining repayment burden and financial risk.

Another key variable, `YTDINTAMT` (Year-to-Date Interest Amount), exhibits multiple dependencies, indicating its influence on the overall financial burden. Additionally, `DISTTCD` (District Code) appears in several connections, signifying the role of regional factors in credit risk assessment.

The network clusters into three primary dependencies: (1) credit limit factors (SBWCLMTAMT, SBTLLMTAMT), (2) installment related attributes (NUMOFINST, INSTA-

Table 5 Conditional Paths in Personal Loan Bayesian Network

Start Variable	Conditional Path & Managerial Insight
PRISECAMT	PRISECAMT → YTDINTAMT → NUMOFINST → INSTALAMT → DPAMT. The principal sanctioned amount establishes the financial foundation of the loan and determines interest accrual, which shapes repayment structure. Managers should consider sanctioned amounts as key drivers of borrower obligations.
DISTTCD	DISTTCD → PRISECAMT → YTDINTAMT → NUMOFINST → INSTALAMT → DPAMT. Regional variation influences sanctioned amounts and interest accumulation, which in turn affect repayment schedules. Managers should adjust lending strategies to local conditions.
BORWORAMT	BORWORAMT → DISTTCD → PRISECAMT → YTDINTAMT → NUMOFINST → INSTALAMT → DPAMT. The initial borrowing request affects sanctioned limits and interest outcomes, guiding repayment terms. Managers should consider initial borrowing intent when structuring loans.
SPLCSAMT	SPLCSAMT → BORWORAMT → DISTTCD → PRISECAMT → YTDINTAMT → NUMOFINST → INSTALAMT → DPAMT. Scheme allocations define borrowing limits and indirectly shape repayment schedules. Managers should align scheme support with borrower capacity to improve portfolio outcomes.
INSAMT	INSAMT → SPLCSAMT → BORWORAMT → DISTTCD → PRISECAMT → YTDINTAMT → NUMOFINST → INSTALAMT → DPAMT. Insurance coverage signals borrower prudence and interacts with scheme allocations to influence repayment. Managers should integrate protection products to enhance repayment stability.
NFMRGAMT	NFMRGAMT → SPLCSAMT → BORWORAMT → DISTTCD → PRISECAMT → YTDINTAMT → NUMOFINST → INSTALAMT → DPAMT. Non-farm margin reflects borrower contribution and affects scheme allocation, borrowing, and repayment flow. Managers should consider margin financing as a signal of borrower stake and repayment discipline.
ORIGINAMT	ORIGINAMT → SPLCSAMT → BORWORAMT → DISTTCD → PRISECAMT → YTDINTAMT → NUMOFINST → INSTALAMT → DPAMT. Original amount represents baseline exposure and shapes sanctioned limits and repayment schedules. Managers should align new credit proposals with historical repayment behavior.
PERIOD	PERIOD → SPLCSAMT → BORWORAMT → DISTTCD → PRISECAMT → YTDINTAMT → NUMOFINST → INSTALAMT → DPAMT. Proposal duration influences scheme allocation, borrowing limits, and repayment terms. Managers should consider duration as a lever for repayment flexibility and risk calibration.
SBTLLMTAMT	SBTLLMTAMT → SPLCSAMT → BORWORAMT → DISTTCD → PRISECAMT → YTDINTAMT → NUMOFINST → INSTALAMT → DPAMT. Subtotal limit aggregates scheme or group limits and affects repayment structure. Managers should ensure group limits align with individual repayment capacity.
RECAMT	RECAMT → SPLCSAMT → BORWORAMT → DISTTCD → PRISECAMT → YTDINTAMT → NUMOFINST → INSTALAMT → DPAMT. Recovery amount reflects past repayment behavior and influences new borrowing and repayment flows. Managers should use recovery history to guide fresh credit exposure.
SBWCLMTAMT	SBWCLMTAMT → SPLCSAMT → BORWORAMT → DISTTCD → PRISECAMT → YTDINTAMT → NUMOFINST → INSTALAMT → DPAMT. Sub-borrower credit limit sets permissible exposure and shapes repayment obligations. Managers should treat credit ceilings as risk control tools.
TPGAMT	TPGAMT → SPLCSAMT → BORWORAMT → DISTTCD → PRISECAMT → YTDINTAMT → NUMOFINST → INSTALAMT → DPAMT. Total proposed general amount defines upper exposure thresholds and downstream repayment. Managers should treat proposal limits as a blueprint for structuring and repayment behavior.
YTDCSUMAMT	YTDCSUMAMT → SPLCSAMT → BORWORAMT → DISTTCD → PRISECAMT → YTDINTAMT → NUMOFINST → INSTALAMT → DPAMT. Year-to-date credit sum constrains borrowing and influences repayment scheduling. Managers should account for cumulative exposure to avoid concentration risk.

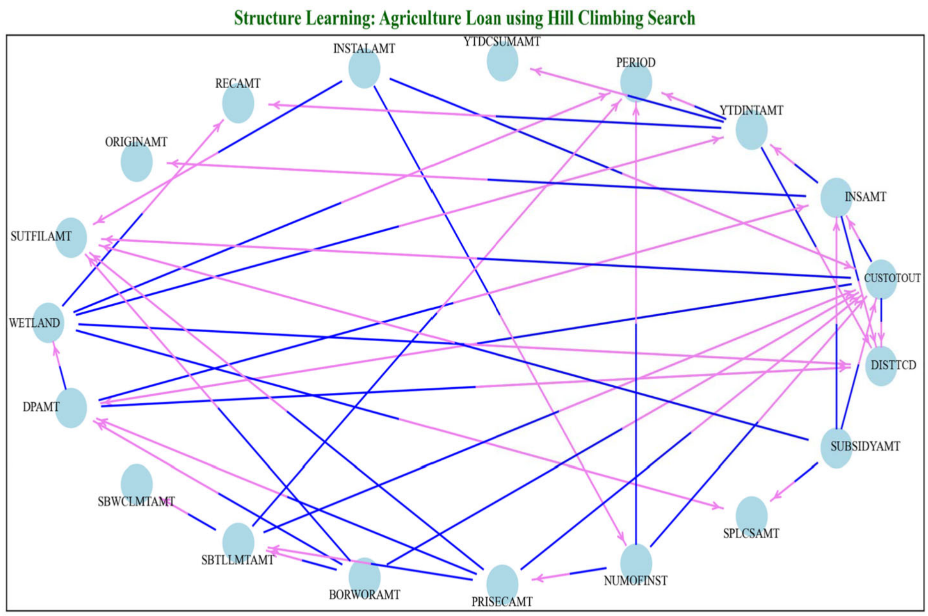


Figure 4 shows the automatically generated Bayesian network for the agricultural loan portfolio. As with Fig. 3, the network output cannot be manually simplified without altering model-learned dependencies. To enhance readability, only the most representative paths are emphasized in the accompanying discussion—particularly those capturing core relationships such as WETLAND → DPAMT → CUSTOTOUT and SPLCSAMT → SUBSIDYAMT → INSTALAMT. These summarize the main interactions between land characteristics, subsidy influence, and repayment structure that dominate the agricultural loan network. Full variable label mappings are provided in Table 12

LAMT), and (3) risk related features (BORWORAMT, CUSTOTOUT). The presence of RECAMT (Recovery Amount) in multiple connections further reinforces its significance in credit risk prediction.

These patterns show that in personal loans, variables such as sanctioned loan amounts (SBWCLMTAMT), installment counts (NUMOFINST), and interest accumulation (YTDINT AMT) are strongly interconnected. These relationships collectively drive repayment obligations and outstanding balances, highlighting the need for advanced risk assessment models that can capture how these factors jointly impact borrower financial stress.

To provide an interpretable overview of these relationships, we summarize the major conditional dependencies in Table 5. Each row shows the starting variable of the path, the complete sequence of connected variables along the dependency chain, and associated managerial insights. This concise tabular allowing readers to quickly understand how individual variables propagate influence across the network and affect credit risk patterns.

5.5.2 *Agricultural loan Bayesian network*

The agricultural loan network shown in Fig. 4 also consists of 19 nodes but has a lower number of edges (40), indicating a less complex structure compared to the personal loan network.

A key differentiating factor is the presence of WETLAND, which plays a significant role in determining financial dependencies. Unlike personal loans, where installment structures dominate, agricultural loans are heavily influenced by land characteristics. WETLAND connects to DPAMT, DISTTCD, INSAMT, and YTDINTAMT, highlighting the strong relationship between land ownership, creditworthiness, and financial risks.

The role of SPLCSAMT (Special Case Subsidy Amount) is another distinguishing feature of the agricultural loan network. Its direct connection to SUBSIDYAMT reinforces the importance of government subsidies in mitigating financial risk. The influence of subsidies on installment amounts (INSTALAMT) and down payments (DPAMT) suggests that policy driven financial support plays a crucial role in agricultural lending.

Regional factors, represented by DISTTCD, exhibit significant connectivity, emphasizing geographical differences in agricultural credit risk. The dependencies around DPAMT suggest that down payments are a primary financial determinant in the agricultural sector, influencing key financial parameters such as CUSTOTOUT, BORWORAMT, and PRISECAMT.

Overall, the agricultural loan network demonstrates a lower level of complexity than personal loans, primarily due to its dependence on land based attributes and government support structures. The presence of fewer connections implies a more straightforward credit risk assessment framework, where risk is dictated by land ownership and subsidies rather than intricate repayment structures.

To provide an interpretable overview of these relationships, we summarize the major conditional dependencies in Table 6.

The comparative analysis in Table 7 reveals differences in network density, influential features, and financial risk factors. Personal loans exhibit higher complexity with a focus on installment based repayment, whereas agricultural loans are significantly influenced by land ownership and government subsidies.

For brevity and readability, only representative dependency paths derived from Figs. 3 and 4 are discussed in detail. The full set of automatically generated connections is available in the supplementary tables.

The heatmaps in Fig. 7 present in A.1 visually validate the structural insights drawn from the Bayesian networks.

5.6 Markov blanket personal loan variable CUSTOTOUT

To uncover the most informative predictors of key variables in our unsupervised framework, we examine the *Markov blanket* of CUSTOTOUT (customer total outstanding) in both personal and agricultural loan contexts.

The Markov blanket of a node in a Bayesian network includes *Direct Causes (Parents):* Variables that directly influence the target, *Direct Effects (Children):* Variables influenced by the target and *Co Causes (Spouses):* Variables that share a child with the target, capturing common upstream influences.

This local structure provides a focused lens on variable interactions without needing the full joint distribution.

Here we analyze the blanket of the central financial risk variable, CUSTOTOUT. In the personal loan network (Fig. 5), the Markov blanket of CUSTOTOUT includes: INSTALAMT, NUMOFINST, DPAMT, YTDINTAMT, PRISECAMT, and RECAMT. These variables represent the minimal set of nodes that render CUSTOTOUT conditionally independent of all others in the Bayesian structure—comprising its direct causes, effects, and co-causes. While variables such as INSAMT (Insurance Amount) or SPLCSAMT (Special Case Subsidy Amount) appear influential in upstream dependency chains (see Table 5), their ef-

Table 6 Conditional Paths in Personal Loan Bayesian Network

Start Variable	Conditional Path & Managerial Insight
NUMOFINST	NUMOFINST → INSTALAMT. This path emphasizes the direct influence of the number of installments on installment amounts. Managers should consider installment counts when structuring repayment schedules.
PRISECAMT	PRISECAMT → NUMOFINST → INSTALAMT. This path emphasizes the central role of installment structuring. The sanctioned amount PRISECAMT indirectly shapes repayment terms via NUMOFINST. Managers should align sanctioned amounts with seasonal repayment ability to reduce stress and improve recovery rates.
SBTLLMTAMT	SBTLLMTAMT → BORWORAMT, SBTLLMTAMT → PRISECAMT → NUMOFINST → INSTALAMT. These two intertwined paths begin with the subordinate loan limit SBTLLMTAMT, which sets the upper threshold of borrowing eligibility. The first path directly influences the proposed borrowing amount BORWORAMT, while the second path flows through the sanctioned amount PRISECAMT, determining the installment structure via NUMOFINST and INSTALAMT. Managers should ensure loan sizing aligns with policy ceilings while structuring repayment schedules.
(DPAMT or SUTFILAMT)	(DPAMT or SUTFILAMT) → CUSTOTOUT → SBTLLMTAMT → PRISECAMT → NUMOFINST → INSTALAMT. This path starts with either the down payment DPAMT or scheme utilization amount SUTFILAMT, which shape net exposure (CUSTOTOUT) and link to borrowing limits. The limit influences sanctioned amounts and installment structures. Managers should treat upfront amounts and scheme fit as levers for controlling credit flow and ensuring repayment viability.
(WETLAND or INSAMT)	(WETLAND or INSAMT) → DPAMT → CUSTOTOUT → SBTLLMTAMT → PRISECAMT → NUMOFINST → INSTALAMT. These nested paths show that agronomic or insurance variables affect initial payments, which propagate through exposure and sanction limits. Managers could use land or insurance attributes to tailor financial obligations to productive capacity or risk profiles.
ORIGINAMT	ORIGINAMT → INSAMT → DPAMT → CUSTOTOUT → SBTLLMTAMT → PRISECAMT → NUMOFINST → INSTALAMT. This path begins with the original loan request ORIGINAMT and flows through insurance and upfront contributions into sanctioned amounts and installment structure. Managers should evaluate how protection mechanisms support lending decisions and affordability.
DISTTCD	DISTTCD → YTDINTAMT → INSAMT → DPAMT → CUSTOTOUT → SBTLLMTAMT → PRISECAMT → NUMOFINST → INSTALAMT. Regional context influences accumulated interest, insurance, and upfront payments, cascading into sanctioned amounts and installment structure. Managers can use location-specific risk profiling to better align credit terms with regional repayment conditions.
(PERIOD or RECAMT)	(PERIOD or RECAMT) → YTDINTAMT → (INSAMT or WETLAND) → DPAMT → CUSTOTOUT → SBTLLMTAMT → PRISECAMT → NUMOFINST → INSTALAMT. These nested paths show that temporal factors or prior recovery amounts set the stage for accumulated burden, which is filtered through insurance or land context, impacting sanctioned amounts and installment planning. Managers should consider past repayment history and timing when designing repayment flows.
SBWCLMTAMT	SBWCLMTAMT → SBTLLMTAMT → PRISECAMT → NUMOFINST → INSTALAMT. This compact path demonstrates a policy-driven flow from working capital limits into subordinate thresholds, influencing sanctioned amounts and installments. Managers should adjust exposure buffers strategically without disrupting repayment alignment.
SPLCSAMT	SPLCSAMT → WETLAND → DPAMT → CUSTOTOUT → SBTLLMTAMT → PRISECAMT → NUMOFINST → INSTALAMT. Starting from scheme-specific support SPLCSAMT, this path includes land context and borrower equity before flowing into sanctioned amounts and installment planning. Managers should leverage scheme eligibility and land type to calibrate loans effectively.
SUTFILAMT	SUTFILAMT → CUSTOTOUT → SBTLLMTAMT → PRISECAMT → NUMOFINST → INSTALAMT. This path shows how subsidy utilization informs exposure and repayment structure. Managers should monitor utilization metrics as real-time signals for customizing repayment design.
YTDCSUMAMT	YTDCSUMAMT → YTDINTAMT → (INSAMT or WETLAND) → DPAMT → CUSTOTOUT → SBTLLMTAMT → PRISECAMT → NUMOFINST → INSTALAMT. These nested paths show that cumulative subsidies and interest influence risk mitigation, borrower stake, and eventual credit configuration. Managers should treat these factors as compound predictors of repayment potential.

Table 7 Comparison of Personal vs. Agricultural Loan Networks

Feature	Personal Loan Network	Agricultural Loan Network
Nodes	18	19
Edges	71	40
Highly Connected Nodes	SBWCLMTAMT, NUMOFINST, YTDINTAMT	WETLAND, SPLCSAMT, DISTTCD
Key Influencers	Credit limit (SBWCLMTAMT), Installment structure (NUMOFINST), Interest (YTDINTAMT)	Land ownership (WETLAND), Government subsidies (SUBSIDYAMT), Region/District (DISTTCD)
Complexity	High Densely connected network with several features having 15+ direct dependencies	Moderate Fewer dense nodes, more localized dependencies
Dependency Patterns	Repayment focused Dependencies center around installment logic and credit burden	Policy and land driven Strong influence from region, land, and subsidy variables
Financial Risk Drivers	Loan burden, accumulated interest, repayment stress (SBWCLMTAMT, NUMOFINST, YTDINTAMT)	Land risk, subsidy dependence, regional exposure (CUSTOTOUT, SUBSIDYAMT, DISTTCD)

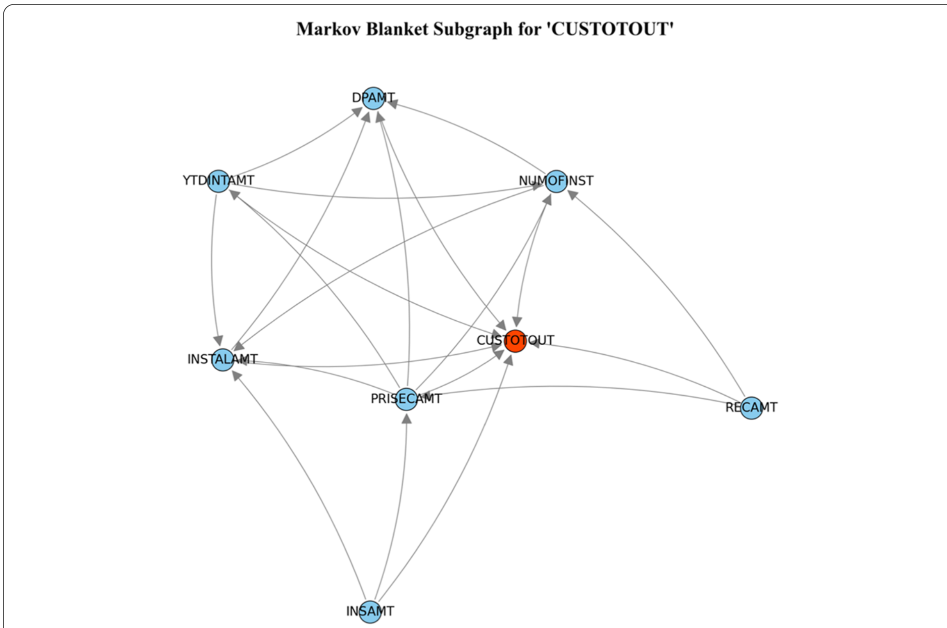


Figure 5 Markov Blanket of CUSTOTOUT – Personal Loan Network

fects on CUSTOTOUT are indirect and mediated through nodes like DPAMT or BORWORAMT. Hence, they are excluded from the immediate Markov boundary according to the learned network. Table 11 further confirms this exclusion, where INSAMT is not marked in the blanket membership row for CUSTOTOUT.

Direct Causes: Variables such as INSTALAMT and DPAMT likely influence the total outstanding amount by controlling the periodic financial burden on the borrower.

Direct Effects: RECAMT may be a consequence of CUSTOTOUT, especially in situations where high outstanding balances trigger collection actions.

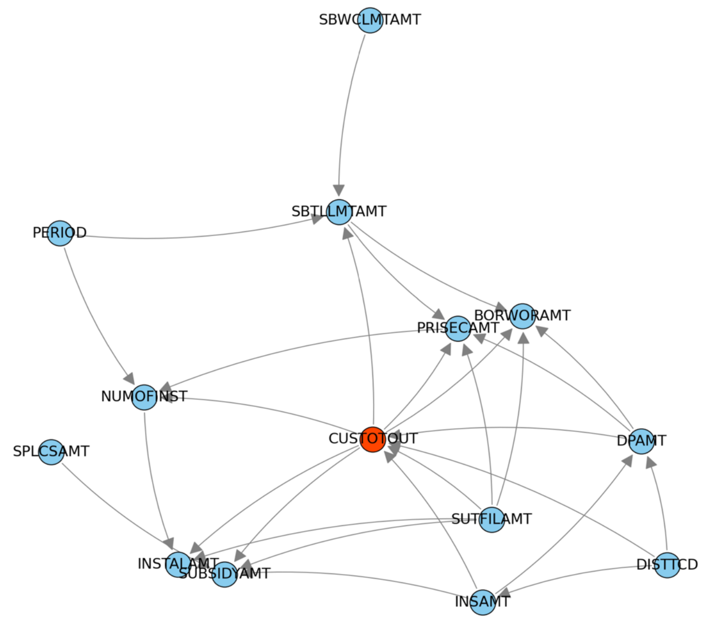


Figure 6 Markov Blanket of CUSTOTOUT – Agricultural Loan Network

Co Causes: Features like PRISECAMT and YTDINTAMT may not directly influence or be influenced by CUSTOTOUT, but are connected through shared relationships with installment patterns or loan tenure.

In the agricultural loan network (Fig. 6), the Markov blanket of CUSTOTOUT includes: DPAMT, SUTFILAMT, WETLAND, INSAMT, SPLCSAMT, DISTTCD, SBTLLMTAMT, PRISECAMT, NUMOFINST, and INSTALAMT. These features jointly capture the immediate land, insurance, and policy factors influencing the customer's total outstanding balance.

Compared with the personal loan blanket, the agricultural loan blanket remains broader in scope, reflecting stronger policy and regional influences. This arises from the inclusion of subsidy-related variables (SPLCSAMT, SUTFILAMT) and the geographic variable (DISTTCD), which jointly shape borrower exposure alongside land and insurance factors.

Direct Causes: Government linked variables like SUBSIDYAMT and WETLAND can influence the outstanding balance, either by reducing liabilities or reflecting credit eligibility.

Direct Effects: Similar to the personal case, INSTALAMT or disbursal elements like SPLCSAMT may reflect how the outstanding amount is broken down or acted upon.

Co Causes: PRISECAMT and NUMOFINST might share common dependencies (e.g., farm size, crop type, or subsidy conditions) that indirectly tie into CUSTOTOUT.

6 Managerial and regulatory interpretations

This section translates the uncovered dependency structures into specific, actionable implications for credit risk management across personal and agricultural loans. By tailoring strategies to the distinct interdependencies observed in each segment, banks can better align operational practices with actual portfolio dynamics.

Table 8 Managerial Implications for Personal Loans

Observed Dependency Insights	Operational / Policy Implication
`SBWCLMTAMT`, `NUMOFINST`, and `INSTALAMT` form strong clusters	Suggests dynamically adjusting installment plans based on sanctioned amounts to manage affordability.
`YTDINTAMT` and `YTDCSUMAMT` are repeatedly linked to risk markers	Calls for frequent tracking of accumulated interest to enable early intervention.
`PRISECAMT` and `BORWORAMT` appear in multiple Markov blankets with repayment variables	Indicates the need for parallel review of primary security values and borrower outstanding balances during risk reassessments.
`PERIOD` and `RECAMT` jointly show up with customer outstandings	Highlights the role of tenure and recovery progress; may justify differentiated restructuring options.
`ORIGINAMT` and `SBTLLMTAMT` connect indirectly via installments and repayment velocity	Underlines the importance of reviewing original sanctioned structures against current limits for realignment.
Multi-feature Markov blankets involving `DISTTCD` suggest regional repayment nuances even in personal portfolios	Points to embedding local market dynamics into loan performance dashboards.

Table 9 Managerial Implications for Agricultural Loans

Observed Dependency Insights	Operational / Policy Implication
`WETLAND` strongly ties to `NUMOFINST` and `DPAMT`	Suggests tailoring installment structures to land profiles, such as longer tenures for wetland plots.
`SPLCSAMT` and `SUBSIDYAMT` frequently influence provisioning and disbursements	Calls for integrating subsidy tracking into regular risk reviews to anticipate credit strain.
`SBTLLMTAMT` and `PROV_TOTAL` appear together with `CUSTOTOUT`	Supports targeted provisioning buffers for specific land types or scheme-driven segments.
`DISTTCD` appears across several blankets with `YTDINTAMT` and `INSTALAMT`	Indicates district-specific policy adjustments, such as varying interest relief thresholds based on local exposure.
`SUTFILAMT` clusters with multiple key repayment and disbursal metrics	Justifies segmenting portfolio dashboards by scheme participation for proactive follow-ups.
Markov blankets are narrower than personal loans, centering on fewer dominant features	Implies simpler, more direct scorecard checks may suffice, freeing resources for deeper monitoring of personal loans.

Personal loans: operational implications

The intricate dependencies among sanctioned limits, installments, interest accumulation, and other repayment related features indicate that personal loans require closely coordinated monitoring across multiple financial dimensions. Table 8 summarizes the key insights.

Agricultural loans: operational implications

In agricultural loans, simpler but more concentrated dependencies around land, provisioning, and subsidy related variables call for a policy focused approach. Table 9 summarizes these insights.

> **Unified Regulatory Insight:** The explicit dependency graphs and Markov blanket analyses offer transparent evidence of how key financial features jointly shape risk across segments. This directly supports internal audit trails and aligns with Basel expectations for documenting credit risk drivers and maintaining clear model explainability, especially when portfolios span diverse loan types.

Table 10 Novel managerial and policy insights from CRAB-Net dependency analysis

Specific Discovered Dependency	Targeted Operational Implication
Personal loan interest (YTDINTAMT) directly conditions RECAMT, revealing that interest build-up accelerates recoveries	Calls for early intervention strategies when year-to-date interest starts to spike, preventing escalation to collections.
ORIGINAMT and PRISECAMT cluster differently around CUSTOTOUT vs. INSTALAMT	Suggests using distinct underwriting logic for setting installment terms versus monitoring outstanding balances.
WETLAND ties closely with SUTFILAMT and PROV_TOTAL in agri networks, showing how subsidies modify provisioning needs	Indicates that collateral and provisioning standards should dynamically adjust based on land type and subsidy utilization.
Markov blankets in agri loans show isolated influence of SPLCSAMT	Points to high sensitivity of agricultural portfolios to single policy changes, requiring scenario analyses around subsidy revisions.
Personal loan blankets exhibit denser multi-feature overlap	Implies stress testing must account for simultaneous shifts in multiple variables, unlike the simpler agri case.

This dual analysis demonstrates how theCRAB-Net framework enables financial institutions to move beyond generic portfolio metrics toward segment specific strategies grounded in detailed probabilistic relationships. It facilitates not only operational improvements but also strengthens compliance with prudential regulations that demand well justified, data driven credit decisions.

To further consolidate the cross-segment managerial and policy implications derived from the dependency structures, we summarize the novel insights in Table 10.

Regulatory Perspective: These insights justify more granular documentation to regulators. For instance, under Basel II and III, banks are expected to document the precise channels through which capital buffers relate to exposure. The CRAB-Net framework provides explicit graphical evidence (such as WETLAND → SUTFILAMT → PROV_TOTAL), going beyond conventional default risk scores. It also supports more targeted stress testing by simulating shocks to interest accumulation in personal loan portfolios versus subsidy dynamics in agricultural portfolios—improving alignment with ICAAP (Internal Capital Adequacy Assessment Process) standards.

6.1 Implications for real-life predictive modelling

The CRAB-Net framework provides actionable insights for operational credit risk modelling. By distinguishing dependency structures across loan types, it supports more targeted feature engineering and validation in predictive systems that follow CRISP-DM or Basel modelling workflows. The network-derived relationships can aid in segment-specific recalibration of scorecards, early-warning systems, and provisioning rules. Integrating such dependency maps into iterative model-risk management and validation cycles can help ensure that predictive tools remain transparent, interpretable, and aligned with supervisory requirements under IFRS 9 and Basel III.

7 Future directions: enhancing the Bayesian network for improved credit risk analysis

Future work can significantly expand the capabilities of this framework by incorporating additional data sources and advanced modeling techniques. One important direction is to enrich the feature space through careful feature engineering, integrating external risk factors that can offer a more holistic view of borrower creditworthiness.

For personal loans, adding macroeconomic indicators such as inflation rates, employment statistics, and general economic sentiment measures can provide valuable context that influences borrowers' repayment abilities. In the agricultural segment, integrating weather data, soil quality metrics, and crop yield forecasts would allow the model to better reflect the operational realities faced by farmers. Incorporating features derived from historical repayment behavior, such as rolling averages of overdue balances or time elapsed since the last missed payment, could also help establish time dependent relationships that capture evolving borrower risk over multiple periods.

Another promising direction is to explore hybrid modeling approaches. Combining Bayesian networks with machine learning methods such as `XGBoost` or LightGBM could help capture more complex, nonlinear interactions among variables that purely probabilistic dependencies might overlook. Additionally, introducing models that explicitly handle temporal sequences, such as Hidden Markov Models or Dynamic Bayesian Networks, would allow for a more dynamic assessment of borrower risk by capturing how financial conditions and repayment patterns develop over time.

The intervention strategies highlighted by this analysis also point toward areas where lending practices can be refined. For personal loans, the strong interconnections among sanctioned loan limits (`SBWCLMTAMT`), installment structures (`NUMOFINST`), and year to date interest burdens (`YTDINTAMT`) suggest that banks could benefit from dynamically adjusting credit limits. Rather than setting static limits, adjusting sanctioned amounts based on a borrower's installment history and capacity could reduce financial stress and improve repayment performance.

In agricultural lending, the observed influence of land characteristics and regional factors, represented by variables such as `WETLAND` and `DISTTCD`, underscores the importance of geographic considerations. The roles of `SPLCSAMT` and `SUBSIDYAMT` indicate that government support has a clear impact on mitigating risk. Banks could therefore refine subsidy allocation by considering regional vulnerabilities, such as exposure to climate risks or historical yield fluctuations. A more targeted subsidy policy could enhance financial resilience and reduce concentrated credit risks in vulnerable areas.

Future work could further enhance the proposed framework by introducing benchmarking mechanisms to evaluate the performance of the unsupervised Bayesian approach. Since `CRAB-Net` operates without predefined labels, conventional supervised metrics are not directly applicable. Instead, techniques such as clustering validation indices or the generation of pseudo-labels could be employed to quantify how effectively the network differentiates among distinct risk profiles. This would provide a more systematic measure of the framework's capability to uncover meaningful credit risk structures.

In addition, future research could incorporate sensitivity and robustness analyses across different temporal windows, economic regimes, or regional subsets. Such investigations would help evaluate the stability of selected features and learned dependencies under varying market contexts. Access to more granular or cross-institutional datasets would further enable validation of the framework's generalizability and adaptability to structural changes in credit behavior.

By integrating these advancements, the `CRAB-Net` approach can develop into a more comprehensive tool for credit risk analysis. These improvements would enhance its ability to predict defaults, support strategic credit decisions, and strengthen compliance with regulatory expectations that emphasize transparent, data driven assessments of financial risk.

8 Conclusion

This study presented CRAB-Net, a framework that integrates row type specific segmentation, SMOTE balancing, XGBoost based supervised feature selection, and unsupervised Bayesian network structure learning to analyze credit risk in heterogeneous banking data. By treating personal and agricultural loans separately, the approach preserved type specific attribute behaviors, allowing the discovery of distinct probabilistic dependency structures through directed Bayesian networks and Markov blanket analysis. These results demonstrate the framework's capacity to reveal key drivers and interdependencies unique to each loan category, enhancing interpretability and supporting differentiated risk management practices. By combining supervised feature refinement with unsupervised exploration of conditional relationships, CRAB-Net provides a transparent tool for credit risk assessment that aligns with regulatory standards such as Basel II and III. Future work will focus on incorporating external macroeconomic or sector specific indicators, exploring temporal or hybrid probabilistic models, and comparing against clustering or sparse graphical techniques to further improve scalability and insight generation in complex financial environments.

Appendix A: Supplementary technical details

A.1 Markov blanket membership matrix (a detailed local dependency structures)

Table 11 presents a binary matrix indicating the presence (✓) of each variable in the Markov blanket of others. A checkmark in cell (i,j) means variable j is in the Markov blanket of variable i. The table lists the immediate set of features most directly influencing or linked to these critical measures, such as customer outstandings and provisioning levels. This detailed mapping supports transparency in how various borrower, account, and loan characteristics jointly shape risk, offering a practical reference for credit risk managers and audit teams.

A.1.1 Managerial insights from local dependency structures

This section distills tailored insights from the Markov blanket analyses of key financial variables. These observations move beyond well known sanctioned limit and installment relationships to reveal subtle multi variable influences that are critical for branch managers and risk teams. By directly connecting these patterns to operational strategies, the analysis bridges the gap between probabilistic dependency models and actionable portfolio management.

Personal loans

Sanctioned limits (SBWCLMTAMT) show dense ties to collateral (PRISECAMT), number of installments (NUMOFINST), and year-to-date interest (YTDINTAMT): Highlights how approving higher limits without corresponding collateral or careful installment planning can elevate future interest burdens, suggesting branches review sanction to collateral ratios.

Table 11 Markov Blanket Membership Matrix

Variable	Type	SBWCLMTAMT	NUMOFINST	PRISECAMT	BORWORAMT	SPLCSAMT	SBTLLMTAMT	YTDINTAMT	DISTTCD	TPGAMT	DPAMT	INSTALAMT	INSAMT	PERIOD	YTDCSUMAMT	RECAMT	ORIGINAMT	NFMRGAMT	WETLAND	SUBSIDYAMT	SUTFILAMT
SBWCLMTAMT	Personal			✓	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓	✓	✓			
	Agri						✓							✓							
NUMOFINST	Personal	✓		✓	✓			✓	✓		✓	✓	✓	✓	✓	✓	✓				
	Agri			✓			✓					✓		✓		✓					✓
PRISECAMT	Personal	✓	✓		✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓			
	Agri		✓				✓				✓			✓							✓
BORWORAMT	Personal	✓	✓	✓		✓	✓	✓	✓	✓			✓	✓	✓	✓	✓	✓			
	Agri						✓				✓										✓
SPLCSAMT	Personal	✓		✓	✓		✓		✓	✓			✓	✓	✓	✓	✓	✓			
	Agri												✓			✓			✓	✓	✓
SBTLLMTAMT	Personal	✓		✓	✓	✓		✓	✓	✓			✓	✓	✓	✓	✓	✓			
	Agri	✓		✓	✓						✓			✓							✓
YTDINTAMT	Personal	✓	✓	✓	✓		✓		✓		✓	✓	✓	✓	✓	✓	✓		✓		
	Agri					✓		✓	✓				✓	✓	✓	✓	✓		✓		
DISTTCD	Personal	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓			
	Agri					✓	✓	✓			✓		✓	✓	✓	✓	✓		✓	✓	✓
TPGAMT	Personal	✓				✓	✓						✓	✓	✓	✓					
	Agri																				
DPAMT	Personal		✓	✓				✓	✓			✓	✓		✓	✓					
	Agri			✓	✓		✓		✓				✓			✓					✓
INSTALAMT	Personal		✓	✓				✓	✓		✓		✓		✓	✓					
	Agri		✓																		✓
INSAMT	Personal	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓			
	Agri					✓		✓	✓		✓						✓			✓	✓
PERIOD	Personal	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓		✓	✓	✓	✓			
	Agri	✓	✓	✓		✓	✓	✓	✓						✓	✓			✓		
YTDCSUMAMT	Personal	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓			
	Agri							✓	✓					✓	✓	✓					
RECAMT	Personal	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓			
	Agri					✓		✓	✓					✓	✓	✓			✓		
ORIGINAMT	Personal	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓		✓			
	Agri							✓	✓				✓								
NFMRGAMT	Personal	✓		✓	✓	✓	✓	✓	✓	✓			✓	✓	✓	✓	✓				
	Agri																				
WETLAND	Personal																				
	Agri					✓		✓	✓					✓		✓					
SUBSIDYAMT	Personal																				
	Agri					✓							✓								✓
SUTFILAMT	Personal																				
	Agri		✓	✓	✓	✓	✓	✓	✓		✓	✓	✓							✓	

Installment counts (`NUMOFINST`) often connect with down payments (`DPAMT`), repayment accumulations (`YTDCSUMAMT`), and recoveries (`RECAMT`): Indicates that structuring more manageable installments may directly influence long term repayment velocity and reduce reliance on recovery processes.

Primary security amounts (`PRISECAMT`) appear closely tied to provisioning flags and interest: Signals that secured loans with higher primary collateral might still warrant provisioning reviews if interest accumulation accelerates, guiding periodic collateral adequacy checks.

Recovery amounts (`RECAMT`) consistently emerge with sanctioned limits and interest: Implies recovery performance is closely related to original sanction decisions and subsequent interest stress, suggesting periodic reevaluation of recovery strategies in high interest or high sanction segments.

Interest accumulation (`YTDINTAMT`) co occurs with provisioning levels and customer outstandings (`CUSTOTOUT`): Points to higher interest accumulation triggering provisioning needs, indicating early identification of accounts with fast growing interest as a key preventive measure.

District code (`DISTTCD`) shows interactions with repayment and collateral factors: Reveals that certain regions might structurally differ in how collateral or installment patterns influence repayments, supporting geo segmented risk strategies.

Outstanding positions (`BORWORAMT`) connect heavily with repayment metrics and sanctioned limits: Suggests that even within sanctioned caps, evolving borrowings may need tighter installment recalibration to mitigate rollover risks.

Agricultural loans

Land classification (`WETLAND`) profoundly shapes repayment structures, appearing with `NUMOFINST` and `DPAMT`: Suggests branch managers could proactively adjust installment planning for different land segments.

Provisioning levels (`SPLCSAMT`) and government subsidies (`SUBSIDYAMT`) often appear together, linked with disbursement (`DPAMT`): Indicates that subsidies might partially offset provisioning pressures, guiding better subsidy targeting to reduce provisioning needs.

Special total limits (`SBTLLMTAMT`) and security values appear with provisioning and repayment patterns: Suggests higher credit ceilings should be carefully aligned with land security values and existing provisioning flags to minimize unexpected defaults.

District code (`DISTTCD`) frequently ties to subsidy and installment structures: Points to region specific repayment behaviors, meaning that regionally tailored subsidy or repayment schemes may stabilize risk.

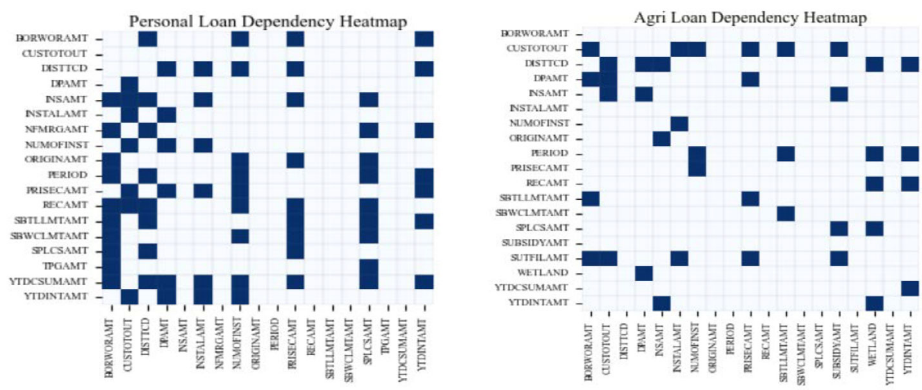
Year-to-date interest (`YTDINTAMT`) connects with provisioning and repayments: Shows accumulating interest directly pushing accounts closer to provisioning triggers, suggesting tighter monitoring of interest build up in agriculture portfolios.

Installment amounts (`INSTALAMT`) tie to provisioning and collateral measures: Suggests loan officers could dynamically revalidate installment appropriateness based on updated provisioning or collateral positions.

Customer outstandings (`CUSTOTOUT`) are influenced by a combined interplay of land type, provisioning, and subsidy flows: Points to a need for integrated dashboards that simultaneously track land, provisioning, and subsidy exposures to inform restructuring decisions.

Security provisioning (`PRISECAMT`) interacts with land and repayment features: Supports adjusting collateral and repayment terms jointly based on land conditions to better protect against sudden defaults.

A.2 Dependency heatmaps



(a) Personal Loan Network (b) Agricultural Loan Network

Figure 7 Dependency Heatmaps for Personal and Agricultural Loan Networks

A.3 Dataset column description

Table 12 Column Description

No.	Name	Description	No.	Name	Description
1	Q	Quarter	42	INCAMT	Interest Not Collected Amount
2	BRCD	Branch Code	43	DISTTCD	District Code
3	CUSTID	Customer ID	44	POPCD	Population Code
4	ACCTID	Account ID	45	UNIFUNFLG	Unit Function Status Flag
5	SEGCD	Segment Code	46	ACCSTACD	Accounting Std Code - Previous Quarter
6	ORGCD	Org. Code	47	FOODNONFLG	Food or Non-Food Sector Flag
7	STFCD	Staff Code	48	RENEWALDT	Date of Renewal
8	RESFLG	Resident Flag	49	ALLCUSTID	Customer ID
9	BKGSINCEDT	Banking Since Date	50	DPAMT	Drawing Power Amount
10	GENCD	Gender Code	51	INSEXPDT	Insurance Expiry Date
11	BIRTHDT	Date of Birth	52	INSAMT	Insurance Amount
12	MARST	Marital Status	53	FULDISFLG	Loan Fully Disbursed?
13	OCUCD	Occupation Code	54	LASCREDT	Last Credit Date
14	DRYLAND	Dry Land Area	55	REPTYPCD	Repayment Type Code
15	WETLAND	Wet Land Area	56	PERIOD	Frequency of Interest Application
16	SBWCLMTAMT	Working Capital Limit	57	YTDINTAMT	YTD Interest Applied
17	PARTBKFLG	Participating Bank Flag	58	YTDCSUMAMT	YTD Credit Summation
18	SBTLLMTAMT	Term Loan Limit	59	QDISMADAMT	YTD Disbursement
19	OPINIONDT	Opinion Date	60	NUMOFINST	Number of Instalments
20	TPGAMT	Third Party Guarantee Amount	61	REPFRQCD	Repayment Frequency Code
21	BORWORAMT	Net Worth of Borrower	62	FIRINSDT	First Instalment Date
22	PRODCODE	Product Code	63	INSTALAMT	Instalment Amount
23	FACCD	Facility Code	64	RECAMT	INCA Recovered Amount
24	SUBFACCD	Sub Facility Code	65	CALC_ASSET	Calculated Asset
25	PRIFLG	Priority Flag	66	TFRDT	Transfer to RA Date
26	DIRFINFLG	Direct Finance Flag	67	REASONCD	Reason for RA Transfer
27	SECTORCD	Sector Code	68	GINAMT	Amount Transferred to RA
28	SCHEMECD	Scheme Code	69	RECALLDT	Recall Date
29	ACTCD	Account ID	70	SUTFILAMT	Suit Filed Amount

Table 12 (*Continued*)

No.	Name	Description	No.	Name	Description
30	SANCTIONDT	Date of Sanction	71	CUSTTYPE	Customer Type
31	SANAUTCD	Sanctioning Authority	72	RETAINAMT	Retainable Claim Amount
32	OPENINGDT	Opening Date	73	WOSACD	Retaining Amount
33	LIMITAMT	Limit Amount	74	CUSTOTLMT	Customer Total Limit
34	DOCREVDT	Document Revival Date	75	CUSTOTOUT	Customer Total Outstanding
35	FOODNONFLG	Food/Non-Food Flag	76	SUBSIDYAMT	Subsidy Amount
36	INTRATE	Interest Rate	77	NFMRGAMT	Cash Security Amount
37	OUTAMT	Outstanding Amount	78	MAR_CALC	March Calculated Asset
38	PRISECCD1	Primary Security Code 1	79	MAR_URIPY	March URIPY
39	PRISECCD2	Primary Security Code 2	80	PROV_TOTAL	Current Provision Amount
40	PRISECAMT	Primary Security Amount	81	IRAC	Asset Classification
41	SPLCSAMT	Specific Collateral Amount			

Acknowledgements
Not applicable.

Author contributions
Minati Rath prepared the main manuscript and analysed data. Hema date reviewed the manuscript.

Funding information
This research received no specific grant from any funding agency in the public, commercial, or not for profit sectors.

Data availability
Data analysed is confidential due to privary protection of bank customers.

Declarations

Ethics approval and consent to participate
Not applicable. This study did not involve any human or animal subjects requiring ethical approval.

Consent for publication
Not applicable. This manuscript does not contain any individual person's data in any form.

Competing interests
The authors declare no competing interests.

Received: 17 July 2025 Accepted: 2 April 2026 Published online: 14 April 2026

References

1. Brown K, Moles P (2014) Credit risk management. Credit Risk Manag, 16
2. Chen C, Lin K, Rudin C, Shaposhnik Y, Wang S, Wang T (2022) A holistic approach to interpretability in financial lending: models, visualizations, and summary-explanations. Decis Support Syst 152, 113647
3. Chen Y, Calabrese R, Martin-Barragan B (2024) Interpretable machine learning for imbalanced credit scoring datasets. Eur J Oper Res 312(1):357–372
4. Bücker M, Szepannek G, Gosiewska A, Biecek P (2022) Transparency, auditability, and explainability of machine learning models in credit scoring. J Oper Res Soc 73(1):70–90
5. Lyócsa Š, Vašaničová P, Misheva BH, Vateha MD (2022) Default or profit scoring credit systems? Evidence from European and US peer-to-peer lending markets. Financ Innov 8(1):32
6. Bellotti T, Crook J (2013) Forecasting and stress testing credit card default using dynamic models. Int J Forecast 29(4):563–574
7. Lessmann S, Baesens B, Seow H-V, Thomas LC (2015) Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research. Eur J Oper Res 247(1):124–136
8. Malekipirbazari M, Aksakalli V (2015) Risk assessment in social lending via random forests. Expert Syst Appl 42(10):4621–4631
9. Baesens B, Van Gestel T, Viaene S, Stepanova M, Suykens J, Vanthienen J (2003) Benchmarking state-of-the-art classification algorithms for credit scoring. J Oper Res Soc 54(6):627–635
10. Pai P-F, Tan Y-S, Hsu M-F (2015) Credit rating analysis by the decision-tree support vector machine with ensemble strategies. Int J Fuzzy Syst 17:521–530
11. Zhang R, Xue L, Wang Q (2023) An ensemble credit scoring model based on logistic regression with heterogeneous balancing and weighting effects. Expert Syst Appl 212, 118732
12. Wang Y, Zhang Y, Lu Y, Yu X (2020) A comparative assessment of credit risk model based on machine learning: a case study of bank loan data. Proc Comput Sci 174:141–149
13. Jagtiani J, Lemieux C (2019) The roles of alternative data and machine learning in fintech lending: evidence from the LendingClub consumer platform. Financ Manag 48(4):1009–1029

14. Altman EI, Marco G, Varetto F (1994) Corporate distress diagnosis: comparisons using linear discriminant analysis and neural networks (the Italian experience). J Bank Finance 18(3):505–529
15. Caruso G, Gattone SA, Fortuna F, Di Battista T (2021) Cluster analysis for mixed data: an application to credit risk evaluation. Socio-Econ Plan Sci 73, 100850
16. Mancisidor RA, Kampffmeyer M, Aas K, Jenssen R (2021) Learning latent representations of bank customers with the variational autoencoder. Expert Syst Appl 164, 114020
17. Leong CK (2016) Credit risk scoring with Bayesian network models. Comput Econ 47(3):423–446
18. Teles G, Rodrigues J, Rabelo RAL, Kozlov SA (2020) Artificial neural network and Bayesian network models for credit risk prediction. J Artif Intell Syst 2(1):118–132
19. Lu J, Wu D, Dong J, Dolgui A (2023) A decision support method for credit risk based on the dynamic Bayesian network. Ind Manag Data Syst 123(12):3053–3079
20. Liu J, Zhang X, Xiong H (2024) Credit risk prediction based on causal machine learning: Bayesian network learning, default inference, and interpretation. J Forecast 43(5):1625–1660
21. Pavlenko T, Chernyak O (2010) Credit risk modeling using Bayesian networks. Int J Intell Syst 25(4):326–344
22. Ballester L, López J, Pavía JM (2023) European systemic credit risk transmission using Bayesian networks. Res Int Bus Finance 65, 101914
23. Pham DT, Ruz GA (2009) Unsupervised training of Bayesian networks for data clustering. Proc Royal Soc A, Math Phys Eng Sci 465(2109):2927–2948
24. Zhang Y, Liu J, Liu Y (2018) Bayesian network structure learning: the two-step clustering-based algorithm. In: Proceedings of the AAAI conference on artificial intelligence, vol 32
25. Gal Y, Ghahramani Z (2016) Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: Proceedings of the international conference on machine learning (ICML)
26. Blundell C, Cornebise J, Kavukcuoglu K, Wierstra D (2015) Weight uncertainty in neural networks. In: Proceedings of the international conference on machine learning (ICML)
27. Kypriotakis G, Cinciripini PM, Versace F (2020) Modeling neuroaffective biomarkers of drug addiction: a Bayesian nonparametric approach using Dirichlet process mixtures. J Neurosci Methods 341, 108753
28. Basel Committee on Banking Supervision (2006) International convergence of capital measurement and capital standards
29. BCBS (2011) Basel III. A global regulatory framework for more resilient banks and banking systems. Bank for International Settlements, Basel
30. Cruz MG, Peters GW, Shevchenko PV (2015) Fundamental aspects of operational risk and insurance analytics: a handbook of operational risk. Wiley, New York
31. Jobst AA (2020) Macroprudential bank solvency stress testing in FSAPs for systemically important financial systems. In: Stress testing. International Monetary Fund
32. Ribeiro MT, Singh S, Guestrin C (2016) "Why should I trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 1135–1144
33. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215
34. Bussmann N, Giudici P, Marinelli D, Papenbrock J (2021) Explainable machine learning in credit risk management. Comput Econ 57(1):203–216
35. Misheva BH, Osterrieder J, Hirsa A, Kulkarni O, Lin SF (2021) Explainable AI in credit risk management. arXiv preprint. arXiv:2103.00949
36. Hurd TR, Gleeson JP (2011) A framework for analyzing contagion in banking networks. arXiv preprint. arXiv:1110.4312
37. Liu Q, Luo Y, Wu S, Zhang Z, Yue X, Jin H, Wang L (2022) RMT-Net: reject-aware multi-task network for modeling missing-not-at-random data in financial credit scoring. IEEE Trans Knowl Data Eng 35(7):7427–7439
38. Kang Y, Chen L, Jia N, Wei W, Deng J, Qian H (2022) A CWGAN-GP-based multi-task learning model for consumer credit scoring. Expert Syst Appl 206, 117650
39. Li Z, Wang X, Yao L, Chen Y, Xu G, Lim E-P (2022) Graph neural network with self-attention and multi-task learning for credit default risk prediction. In: International conference on web information systems engineering, pp 616–629
40. Rath M (2023) Adaptive modelling approach for row-type dependent predictive analysis (RTDPA): a framework for designing machine learning models for credit risk analysis in the banking sector. arXiv preprint. arXiv:2311.10799
41. Rath M, Date H (2025) Quantum powered credit risk assessment: a novel approach using hybrid quantum-classical deep neural networks for row-type dependent predictive analysis. EPJ Quantum Technol 12(1):75
42. Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, Wirth R (2000) CRISP-DM 1.0: Step-by-step Data Mining Guide. SPSS Inc
43. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357

Publisher's note