Bayesian Networks Improve Out-of-Distribution Calibration for Agribusiness Delinquency Risk Assessment

Ana C. Teixeira[*][†]
ana.teixeira@traivefinance.com
Traive Inc.
Brookline, Massachusetts, USA

Hamed Yazdanpanah[†]
hamed.yazdanpanah@traivefinance.com
Traive Inc.
Brookline, Massachusetts, USA

Aline O. Pezente
aline.oliveira@traivefinance.com
Traive Inc.
Brookline, Massachusetts, USA

Mohammad M. Ghassemi
mohammad.ghassemi@traivefinance.com
Traive Inc.
Brookline, Massachusetts, USA

ABSTRACT

Automated credit risk assessment plays an important role in agricultural lending. However, credit risk assessment in the agricultural domain has unique challenges due to the impact of weather, pest outbreaks, commodities market dynamics, and other volatile forces that drive risk. Training a model to account for these factors requires immense data assets that are challenging to obtain. Indeed, even the best credit risk assessment models in this domain are trained using data from single-institutions that often focus on dedicated geographical regions, or singular commodities. Hence, most agricultural credit risk models exhibit poor out-of-domain performance. In this paper, we use a novel dataset describing nearly 100 thousand historical loans, sourced from 9 large agricultural lenders to train a Bayesian network model for loan delinquency classification. The proposed model exhibited improved calibration (relative improvement in Expected Calibration Error) in out-of-domain performance tests when compared to three state-of-the-art credit risk scoring approaches: Logistic regression (81 ± 15% improvement), XGBoost (80 ± 14% improvement), and an Artificial Neural Networks (7 ± 2% improvement). We conclude that Bayesian networks provide better modeling of agricultural credit risk by combining (limited) data assets with expert domain knowledge. Our approach is likely to generalize to any credit risk assessment task where small sample sizes is of concern.

KEYWORDS

Bayesian network model, out-of-distribution generalization, probability of delinquency, credit risk assessment, agricultural loan

ACM Reference Format:
Ana C. Teixeira, Hamed Yazdanpanah, Aline O. Pezente, and Mohammad M. Ghassemi. 2018. Bayesian Networks Improve Out-of-Distribution Calibration for Agribusiness Delinquency Risk Assessment. In *Proceedings of 4th ACM International Conference on AI in Finance (ICAIF '23)*. ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Agriculture is critical to food security, and thus, global stability and propriety. There is an increasing investment by public and private financial institutions, such as commercial banks, development banks, investment funds, and industry suppliers, on food production and agriculture industry [8]. The consequences of an incorrect investments are not only problematic for the lender, but also the global food supply, and especially vulnerable populations (which are more sensitive to fluctuations in commodities prices).

Given it's importance, the private sources that finance agricultural activity need robust ways to assess the risk of their investments. The current method are the use of human analysts that leverage a combination of strong domain knowledge, with some data (e.g. from credit bureaus) to assess the credit worthiness of farmers and agricultural businesses seeking loans. However, despite the expertise of many credit analysts, the current method is prone to errors because (1) there are no global standards for how agricultural credit risk assessment aught to be performed and (2) existing analytic tools to support the analysts (such as consumer credit risk scores) are not appropriate because they were not purpose-built for the agriculture; it's not reasonable to assume that a consumer credit risk score can provide a comprehensive assessment of risk because many of the factors that impact agricultural risk are completely beyond the control of the farmer (weather, global conflicts, etc.).

Clearly, there is a need for access to better data and, if this data could be collected, machine learning can enable more efficient and robust risk assessment in the agricultural domain. However, the intricacies of data in the Agriculture industry pose considerable challenges to collect, at scale. Agriculture, by its nature, is a sector heavily influenced by a multitude of uncontrollable and hard to predict factors, such as weather patterns, pest outbreaks, macroeconomic events, geopolitical conflicts, and commodity price volatility. These factors introduce high variance and irregularities into the data that traditional models struggle to capture effectively. Additionally, the vast temporal and spatial differences in agricultural

[*] Corresponding author
[†] These authors contributed equally.

ICAIF '23, November 27–29, 2023, New York, NY
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

practices across regions, coupled with infrequent and inconsistent data collection methods, results in a high degree of data sparsity and lack of homogeneity. The temporal aspect of agricultural data also introduces seasonality effects that are difficult to account for. Moreover, the industry is impacted by intricate policy changes and market dynamics, furthering the already high dimensional and complex nature of the data. Taken together, the complexities result in data assets that are small, incomplete, which leads to machine learning models that fail to generalize effectively, which then reinforces the need for the involvement of human experts. Breaking this cycle will require the application of methods that can use a combination of limited data and expert knowledge, to provide robust assessments of risk in situations beyond their original scope of learning.

1.1 Related Works

Automated credit risk assessment using machine learning presents a promising approach to resolving unique credit risk complexities within the agricultural ecosystem. Widespread applications of AI and machine learning in finance, underscored by an extensive body of literature on these methods in credit risk assessment [1, 6, 20, 23], support that these methodologies can be effectively adapted to the nuances of agricultural credit. Within agricultural finance, there has been an increasing interest in harnessing the power of machine learning to address credit risk, indicating the willingness of the field to adopt innovative, data-driven strategies [2, 5, 11, 19, 25]. Despite these strides, a review of related work reveals that no existing studies have examined the *out-of-distribution* generalization capabilities of their proposed models for estimating credit risk. Hence, while the existing work provides crucial groundwork, there is a clear opportunity to advance the field by developing models with reliable out-of-distribution performance.

Frequentist machine learning models, despite their extensive use in credit risk assessment, struggle with accounting for uncertainty and generalizing to new data [22]. On the other hand, previous research has highlighted Bayesian networks as a promising alternative in credit risk assessment, specifically for their capacity to encapsulate inherent uncertainty in the data through the use of prior distributions [14, 15, 17]. These networks offer more than point predictions, extending to quantify the uncertainty or confidence around these predictions. This attribute becomes particularly advantageous when dealing with data not represented in the training set. Furthermore, the Bayesian networks' graphical representation enables experts to understand and validate the model's learning process. The capabilities of Bayesian networks can potentially enhance the model's robustness to out-of-distribution generalization, when the model development process is thoughtfully designed. Our research provides a contribution to the field by demonstrating that Bayesian networks achieve enhanced out-of-distribution performance on credit risk assessment in agricultural finance.

Bayesian networks serve as a suitable framework for implementing methods aimed at improving out-of-distribution generalization, a major challenge in artificial intelligence systems [18, 24]. Bayesian networks, by facilitating causal representation learning, are at the forefront of methods that effectively counter spurious correlations by accounting for latent variables. Additionally, they can address data shift, given their structural capacity to integrate shift axes as features. Consequently, the impact of these shifts on the target variable can be reflected in the prior distribution, a crucial step toward model robustness against underlying distribution changes. These capabilities position Bayesian networks as a core instrument for principled model design when targeting out-of-distribution generalization. Our work, building on this theoretical foundation, offers empirical substantiation for the efficacy of this approach.

In the quest for robust credit risk models, the emphasis on calibration is paramount. Indeed, in AI's application to finance, well-calibrated classification models are arguably of greater relevance to credit risk assessment than those of excellent classification performance alone [3, 4, 21]. Calibration is important for risk assessment because it measures the reliability of the model's predictions [7, 16]; hence, by accurately measuring uncertainty, well-calibrated models contribute to efficient portfolio management and improved financial performance.

In summary, while machine learning techniques, particularly Bayesian networks, hold great promise in credit risk assessment, their full potential is yet to be explored. Addressing out-of-distribution generalization, as measured by calibration, will lead to more robust and reliable credit risk models. Our research represents a step in this direction.

1.2 Contributions

This paper introduces a novel application of a Bayesian network to overcome challenges in credit risk assessment in agricultural finance, especially in out-of-distribution scenarios. Notably, Bayesian networks incorporate prior knowledge, a valuable trait when data is limited or of low quality, as is frequently the case in the agricultural credit risk domain [15]. They serve as robust mechanisms against overfitting, with their capacity to assimilate domain expertise contributing significantly to model resilience when tackling out-of-distribution inferences.

A key contribution of our work is the demonstration of the Bayesian network's superior out-of-distribution performance, as gauged by calibration - the most relevant metric in this context. We adopted the Bayesian network to predict agricultural loan delinquency, considering the sector's characteristics such as small datasets, low-quality data, and the necessity to incorporate domain knowledge. This approach was shaped through an in-depth collaboration with domain experts, resulting in a model that encapsulates the complexities of agricultural credit risk knowledge effectively and provides calibrated estimates for robust out-of-distribution predictions.

The paper is organized as follows: Section 2.1 details the dataset, feature definitions, strategy for handling missing values, and the discretization of continuous variables. The specifics of the proposed Bayesian network are discussed in Section 2.4. The results, including the calibration-based performance comparison with other frequentist models, are presented in Section 3. The discussion is presented in Section 4. Finally, Section 5 provides conclusions and potential directions for future research.

2 METHODS

2.1 Dataset

In this study, we utilize a dataset comprised of 97,235 agricultural loans granted to 31,900, Brazilian farmers sourced from nine of the largest financial and supply chain institutions in Brazil; the nine institutions are anonymized and shown by capital letters from A to I in this study. Due to differing credit policies, the characteristics of the data was significantly different across the nine institutions. The loans were typically issued at the onset of the farming season and repayment was expected to be fully executed by the end of the crop season, upon harvest and sales. The target variable in this dataset is a dichotomous variable indicating if the loan was unsuccessfully repaid within 90 days of the loan's due date (i.e., a delinquency event was coded as 1). Detailed characteristics of each institution, including the number of loans issued, the count of unique borrowers, the time range of the data, and the respective delinquency rates, are presented in Table 1. Notably, the number of loans sources across across the institutions varied significantly, ranging from a minimum of 1,019 loans (see row I, Table 1) to a maximum of 31,095 loans (see row A, Table 1). More importantly, there is a significant variation in delinquency rates among these institutions, ranging from as low as $1.05\%(InstitutionA)toashighas22.27\%$ (Institution F). This variability highlights the disparate credit policies across these institutions emphasizing the the need for a modeling framework that can effectively generalize out-of-distribution.

2.2 Features

Each loan in our dataset was characterized by a total of nine features; four of the features were the output of separate models ("Scores" with values ranging from 0 to 1000) designed to characterize key contributors to risk, while the remaining five features are descriptive of the farmer and their farm. We provide additional details about the nine features below.

2.2.1 *Scores (n = 4):*

(1) Agronomic Score: the Agronomic Score is the output of an ML Model that uses historical agronomic and weather patterns to predict the expected yield per hectare for a given farmer, at a given location, in a given season. A higher yield results in a higher agronomic score. We expect this feature to be useful because of the correlation between crop yield and the farmer's income - which will be the source for loan repayment.

(2) Market Score: The market score is the output of an ML model that uses historical commodities sales price, the crop portfolio, production/logistics costs, and expected yield to predict the farmer's operational profits. A higher farmer income results in a higher market score. We expect this features to be useful because a higher operational margin allows for more cash available to facilitate loan repayment.

(3) Financial Score: The financial score is the ratio of: the farmer's outstanding indebtedness , in comparison to his/her expected profit. A higher ratio results in a higher financial score. We expect this feature to be useful because it reflects the proceeds available for the repayment of the loan; when this feature is closer to zero, it means that a more significant part of the farmer's profit is used to pay current debts, and less profit will remain to repay the requested loan on the due date.

(4) Behavior Score: The behavior score is a consumer credit risk score sourced from a credit bureau that is fine-tuned for use in agricultural finance. A higher consumer credit score results in a higher behavior score. Although most of the farming activity signals and farmers' financial life are not captured by credit bureau data, they have valuable information about farmers' consumer behavior, such as paying bills. A good credit bureau score shows that the farmer has good behavior regarding paying bills, even if not related to his/her business. Delay or delinquency on the payment of consumer loans and bills is an early alarm of financial stress or irresponsibility; sooner or later, it will impact his/her farming business.

Table 1: Dataset description; our dataset comprised of 97,235 agricultural loans granted to 31,900, Brazilian farmers sourced from nine of the largest financial and supply chain entities in Brazil (rows A - I).

Institution	# samples	# farmers	Time range	Delinquency (%)
A	31,095	8,041	2013-2021	1.05
B	25,819	11,613	2013-2018	2.52
C	23,877	6,792	2014-2022	2.07
D	5,391	1,246	2013-2022	5.51
E	4,127	908	2013-2022	1.43
F	2,685	1,090	2013-2020	22.27
G	1,689	1,357	2019-2022	1.66
H	1,560	476	2013-2022	14.42
I	1,019	377	2014-2022	13.94
Total	97,235	31,900	2013-2023	2.90

2.2.2 *Farm and Farmer Characteristics (n=5):*

(1) Ratio of Short-Term Debt to Total Planted Area: This feature represents the ratio of the farmer's outstanding short-term debt, to their total planted area (RSTDTPA). Short-term debt is defined as all debts that must be paid in the next twelve months. Note that, as mentioned earlier, all loans in the dataset have a duration of crop season. Thus, short-term debt includes all debts that the farmer should pay before the due date of the loan.

(2) Ratio of Long-Term Debt to the Farm Area: This feature represents the ratio of the farmer's long-term debt over their total farm area (RLTDFA). Long-term debt is all debts that must be paid in a period longer than one crop season. Long-term debts are generally used for investments in the farm, such as machinery renewal, installing irrigation systems, etc.

(3) Land Lease Costs: This feature represents the cumulative costs paid by the farmer in the crop season to rent the land from a third party; if the farmer owned their land, this value would be 0.

(4) Credit History: This feature represents the credit history of the farmer with the institution from which they are requesting the loan; the feature may take one of three categories

reflecting if the farmer's last loan was: successfully paid, delinquent, or if the farmer was a new client in the institution's portfolio.

(5) Main Crop: This features represents the main crop grown by the farmer; it has eight categories: soybean, summer corn, winter corn, wheat, rice, Arabic coffee, robusta coffee, sugarcane. Note that the farmer may plant various crops during the loan's life cycle, whereas the one with the most significant area is considered as the main crop.

The first, second, and thrid quartiles (Q1, Q2, Q3), and the percentage of data that were missing for each continuous feature is described in Table 2. As can be seen, many of the features were missing for a significant fraction of the loans (e.g. the behavior score was missing for 95.9% of loans). This is consistent with our expectations when dealing with agricultural loans, sourced from multiple institutions. For a given loan, we addressed missing values by assigning the median value for any missing features.

Regarding the categorical variables, there were no missing values. For the credit history feature: 73,621 samples paid their last loan successfully, 1,954 samples were delinquent on the last loan, and 21,660 samples were new farmers in the intuition's portfolio. For the main cropt feature: 66.3% of the samples were soybean, 21.3% were sugarcane, and the remaining were summer corn, Arabic coffee, rice, wheat, robusta coffee, and winter corn in decreasing order of sample size.

2.3 Assessment of Concept Drift

We assessed the data for concept drift (change in the distribution of the data) across four axes: institutions, crop, state, and year. Concept drift was assessed using the Jensen-Shannon Divergence (JSD) test.

Suppose that $\mathcal{A}$ = {Institution, State, Crop, Year}. Then, for a given feature f, define the following set of JSDs,

$$\mathcal{J}_f = \bigcup_{a\in\mathcal{A}} \left\{JSD_{1\le i<j\le \#a}(\mathbf{f}_{a_i}, \mathbf{f}_{a_j})\right\},$$

where $\mathbf{f}_{a_i}$ is the vector of feature values f for all samples belong to axis a_i, and $\#a$ denotes the number of unique values in the axis a. Then, the 90^{th} percentile of $\mathcal{J}_f$ is computed by

$$P_{90}(f) = \text{percentile}(\mathcal{J}_f, 90).$$

Table 2: The first, second, and third quartiles of the continuous features in our data, and their missing value percentages. See 2.2 for a description of features.

Institution	Q1	Q2	Q3	Missing (%)
Agronomic score	366	500	625	62.2
Market score	275	377	550	62.2
Financial score	650	850	950	97.4
Behavior score	806	831	852	95.9
RSTDTPA	0	17	235	71.3
RLTDFA	0	0	157	41.1
Land Lease Costs	0	0	50900	81.2

Table 3: Analysis of concept drift across various features using Jensen-Shannon divergence. The table quantifies how much the distribution changes when samples' characteristics vary along four axes: institution, crop, state, and year. Each cell indicates the number of Jensen-Shannon divergence values above the 90th percentile corresponding to each axis-feature pair. Most data drift occurs in the onstitution axis, suggesting significant shifts in data distribution of samples across institutions.

Feature	Axis			
	Institution	Crop	State	Year
Agronomic score	8	2	5	0
Market score	6	4	4	1
Financial score	11	2	0	0
Behavior score	3	10	2	0
RSTDTPA	6	0	9	0
RLTDFA	4	2	7	1
Land Lease Costs	0	5	6	0
Main crop	7	0	5	0
Credit history	14	0	0	0
Sum	59	25	38	2

Subsequently, we determine the total number of JSD values above the 90^{th} percentile are associated with each axis; the count is denoted by

$$\text{Count}(f, a) = \sum_{j\in\mathcal{J}_{f,a}} \mathbb{1}(j, P_{90}(f)),$$

where $\mathbb{1}$ stands for the indicator function, and $\mathcal{J}_{f,a}$ is the subset of $\mathcal{J}_f$ in which their axis is a. These values for all features and axes are reported in Table 3. As can be seen, most data drift happened along the institution axis. In other words, the greatest shifts in the data distribution occurred when the institution was changed.

Since most data drift is observed when the institution is changed among the samples; thus, the label shift is also analyzed inside each institution. To this end, the Mann-Whitney U test is employed. For all samples inside each institution, they are divided into two subsets of successful and unsuccessful loan payments. Then, the Mann-Whitney U test is computed for all continuous features in these two subsets, and their p-values are reported in Table 4. It can be seen in this table that, for all institutions, there are evident label shifts when analyzing RSTDTPA and financial score features. When an entry is denoted by NA in this table, it means that for the corresponding feature and institution, all samples in the payment and delinquency subsets were identical, and the Mann-Whitney U test cannot be computed; this was true for institutions D, G, H, and I, where the land lease cost was zero.

Also, since the Mann-Whitney U test is only calculable for numerical variables, the χ^2 test is adopted for analyzing the label shift in categorical variables. The results are summarized in Table 5. As can be observed, the label shift is noticeable for all institutions when evaluating credit history and main crop features.

Table 4: Inter-institutional label shift analysis using Mann-Whitney U test: significant changes detected in continuous features.

Feature	Institution									
	A	B	C	D	E	F	G	H	I	All
Agronomic score	0.70	0.91	0.78	0.99	0.99	0.99	0.93	0.99	0.99	1
Market score	0.38	0.99	0.26	0.006	0.95	4×10^{-11}	0.96	0.98	0.99	0.98
Finnacial score	4×10^{-17}	9×10^{-5}	3×10^{-30}	8×10^{-3}	3×10^{-3}	2×10^{-8}	0.01	2×10^{-4}	2×10^{-23}	1×10^{-26}
Behavior score	2×10^{-5}	0.75	0.02	0.04	0.84	0.57	0.43	0.63	0.39	2×10^{-4}
RSTDTPA	2×10^{-13}	1×10^{-7}	1×10^{-58}	4×10^{-15}	6×10^{-12}	5×10^{-10}	0.04	1×10^{-19}	2×10^{-30}	2×10^{-124}
RLTDFA	0.002	0.003	1×10^{-10}	0.20	0.84	NA	0.12	2×10^{-6}	7×10^{-11}	1×10^{-7}
Land Lease Costs	3×10^{-30}	3×10^{-4}	2×10^{-40}	NA	3×10^{-7}	0.77	NA	NA	NA	3×10^{-4}

Table 5: Categorical feature label shift analysis by institution using Chi-Squared test: notable differences in 'Credit History' and 'Main Crop' across institutions.

Feature	Institution									
	A	B	C	D	E	F	G	H	I	All
Credit history	0	0	0	0	0	0	0	6×10^{-311}	5×10^{-118}	0
Main crop	0	2×10^{-129}	0	8×10^{-44}	0	8×10^{-15}	2×10^{-7}	5×10^{-35}	0.02	0

2.4 Proposed Approach

In this section, we describe the method used to assess delinquency risk in agricultural loans. As demonstrated in the previous section, the data in the agricultural finance sector suffers from several quality issues: sample sizes are relatively small (see Table 1), many features are missing values (Table 2) and the data distributions between different institutions change significantly (Table 3). To account for these data quality issues, human credit analysts will rely heavily on "experts" to provide *apriori* beliefs about the importance of the features. These limitations of the data, and the existence of strong priors from experts position Bayesian networks an excellent fit for modeling risk in this domain.

A Bayesian network is a probabilistic graph model which illustrates a set of variables (nodes) and their conditional relationship and dependencies through a directed acyclic graph (DAG) [9, 10]. Each node in the DAG stands for a feature, and the edges between them define probabilistic relationships linking the corresponding features. Dependencies between features have different strengths, and they are measured by conditional probability distributions. Bayesian networks are capable of integrating observed data with prior knowledge to improve predictions under uncertainty. They bring forth an explicit and natural perception of the dependencies among features; thus, they permit comprehensible interpretation and causal reasoning.

The Bayesian network employed in this study is represented in Figure 1, where the credit performance node is the model target. Expert knowledge in the agricultural credit risk domain is extensively utilized in constructing the network to imitate credit analysts' structure of thinking when assessing an agricultural loan. This expertise helps in defining the nodes (variables) of the network and their relationships based on a deep understanding of the actual conditions influencing credit delinquency in agribusiness. Therefore, the proposed network structure is a close representation of the real-world intricacies of credit delinquency scenarios in agricultural loans.

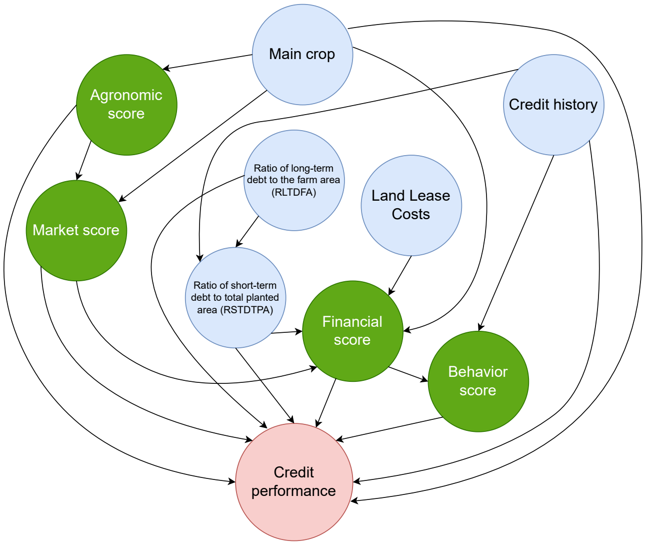

Figure 1: The proposed Bayesian network for agricultural loan credit risk assessment.

In the agricultural loan evaluation, the crop type is the origin of the assessment. It has an immediate impact on the agronomic score (expected crop yield production), market score (expected operational profit), financial score (the ratio between the debts and profit), and credit performance. Thus, there are edges between the main crop node and the nodes corresponding to the mentioned features in the network. Also, there is a direct edge between the agronomic score and market score because higher (lower) expected crop yield production implies higher (lower) expected operational profit. Furthermore, there is an edge connecting the agronomic score and credit performance nodes. The market score is linked to the credit performance as well since the farmer's operational profit amount is directly related to the loan repayment capacity. Additionally, it is connected to the financial score, as the financial score is the ratio between the farmer's debt and profit.

The RLTDFA node is linked to the RSTDTPA and credit performance nodes. Particularly, it is connected to the RSTDTPA because some installments of the long-term debts must be paid within the following twelve months of the loan request date, and they should be considered short-term debts. Note that it does not need to be directly connected to the financial score since it impacts this score indirectly via the RSTDTPA. Besides the main crop node that impacts four nodes directly, the credit history is another important node that impacts three nodes immediately. It is connected to the RSTDTPA because if the farmer did not pay the last loan, he/she is carrying the debts of the previous season to the current season, which should be paid in the following twelve months. Also, it is linked to the behavior score since this score is a measure of the farmer's behavior on consumer loans and daily bill payments. Indeed, when the last season loan has not been paid successfully, it is expected that some consumer loans or bills have defaulted because they were less critical from the farmer's perspective and not related to his/her business. In addition, this node is directly connected to the model target since the last loan performance is an excellent indicator of the subsequent loan output, as observed in the label shift analysis in Table 5.

There is an edge between the RSTDTPA and financial score nodes because the short-term debt is used in the financial score computation. Moreover, the RSTDTPA is directly connected to the credit performance, and it can be seen in Table 4 that this feature has a different distribution between successfully paid and delinquency samples in all institutions. The land lease costs node is only linked to the financial score node since the land lease costs is similar to the short-term debt (it should be paid during the season) and is employed in the financial score computation. The financial score node is connected to the behavior score and credit performance nodes. This feature is a measure of the farmer's indebtedness and describes the farmer's capacity for successful payments. Therefore, besides the credit performance, it impacts the credit bureau data and, consequently, the behavior score. Finally, the behavior score node is only linked to the credit performance node since negative records in the consumer loan and bill payments are early alarms of financial stress, which will impact the farmer's business.

2.4.1 Discretization Approach. To prepare the data for use in a discrete Baysian Network, all continuous features were first discretized. the bins used for the discritization were selected by two credit analysts with over 20 years of collective experience in agribusiness. The thresholds were defined so that when a feature was shifted from one category to another, a significant change in the downsteam risk attributable to that feature was expected by the expert. Finally, the prior conditional probability distribution of the credit performance node was defined and validated by the experts. The priors were used to manage the heterogeneity of data drawn from population-wide information and expert assumptions about agricultural credit risk. These priors enabled us to incorporate additional knowledge into the proposed model, filling in the gaps introduced by heterogeneous data. This integration was intended to improve the model's robustness and overall performance in the out-of-distribution generalization.

2.5 Baselines

Following a review of the literature, three baseline models were selected: eXtreme Gradient Boosting (XGBoost) [13], L2 penalized Logistic Regression (LR) [12] and Artificial Neural Networks (ANN) [20]. The selected baselines represent the best-performing modeling frameworks reported in the systematic review by *Shi et al.* [20]. All model hyperparameters were selected using Bayesian optimization.

2.5.1 Metrics for the Comparison. To compare the models in this work, calibration metrics, such as Expected Calibration Error (ECE), Average Calibration Error (ACE), Maximum Calibration Error (MCE), and Brier score, are reported.

3 RESULTS

In this section, the dataset described in Section 2.1 is utilized to predict delinquency events (unsuccessfully repayment within 90 days of the loan's due date). Thus, the target for the model was a dichotomous variable where the delinquency was coded as 1 and repayment as 0. The model's input has nine features, as explained in Section 2.1. First, the missing values in the dataset are imputed by assigning the median value for any missing features. Then, the continuous features are converted to discrete variables. Various methods can be employed in feature discretization, such as Jenks natural breaks optimization, percentiles, and expert-based discretization. In this study, expert knowledge is used to discretize the feature values, as described in Section 2.4.1. Thus, the bin thresholds for each feature are defined according to expert recommendations. Furthermore, the prior conditional probability distribution of the delinquency event (credit performance) node of the Bayesian network is provided by experts in agricultural loans.

After the imputation and discretization, the Bayesian network described in Section 2.4 is employed to train the model. Among four possible axes of data drift (Institution, Crop, State, and Year), it is observed in Table 3 that the most data drift happened when the institution changed in the dataset. Since the purpose of this study is to show the out-of-distribution generalizability of the proposed Bayesian network, this model is trained on all institutions except one and is tested on the held-out institution. This experiment is repeated so that all institutions are used one time as the test set. Table 6 reports the ECE, ACE, MCE, Brier score values for the Bayesian Network (BN), ANN, XGBoost, and LR models when each institution held out as the test set. Note that all metric values vary between 0 and 1, and a value closer to zero means that the model is more calibrated.

According to each metric, the number of institutions in which the each model has a superior performance to other models is reported in Table 7. It can be observed that, according to all metrics, the Bayesian networks attained the highest performance on the most number of institutions. After the Bayesian network, on the second place, the ANN obtained the best performance on some institutions. For no metric and no institution the XGBoost obtained the best performance, and it is the worst model regarding the calibration

Table 6: ECE, ACE, MCE, and Brier metric values of the Bayesian network, ANN, XGBoost, and LR models on the different held-out institutions.

Metric	Model	Institution								
		A	B	C	D	E	F	G	H	I
ECE	BN	**0.066**	0.020	**0.063**	**0.033**	**0.061**	**0.146**	0.076	**0.115**	**0.073**
	ANN	0.068	**0.014**	0.070	0.038	0.063	0.198	**0.071**	0.137	0.132
	XGBoost	0.284	0.485	0.291	0.549	0.482	0.260	0.553	0.431	0.415
	LR	0.489	0.452	0.479	0.445	0.486	0.263	0.483	0.358	0.361
ACE	BN	**0.280**	**0.270**	0.232	**0.138**	0.194	**0.098**	0.202	0.139	**0.075**
	ANN	0.284	0.491	**0.230**	0.145	**0.185**	0.198	**0.195**	**0.117**	0.112
	XGBoost	0.426	0.491	0.436	0.522	0.512	0.284	0.393	0.498	0.419
	LR	0.495	0.682	0.489	0.458	0.493	0.257	0.491	0.451	0.428
MCE	BN	0.666	**0.666**	**0.714**	**0.411**	0.473	**0.208**	**0.497**	**0.466**	0.181
	ANN	0.650	0.989	0.730	0.428	**0.443**	0.249	0.512	0.517	**0.152**
	XGBoost	0.701	0.683	0.720	0.907	0.696	0.353	0.619	0.904	0.619
	LR	**0.500**	0.972	0.741	0.475	0.500	0.519	0.500	0.739	0.500
Brier	BN	0.025	**0.029**	**0.032**	**0.051**	0.028	**0.193**	0.026	**0.133**	**0.094**
	ANN	**0.021**	0.035	0.034	0.058	**0.026**	0.224	**0.025**	0.141	0.138
	XGBoost	0.111	0.271	0.130	0.353	0.252	0.226	0.325	0.305	0.325
	LR	0.250	0.230	0.250	0.250	0.250	0.240	0.250	0.251	0.250

metrics. The LR model only achieved the highest MCE on one institution. Furthermore, for each metric in this table, it is shown what percentage of the total population in the dataset is composed by the institutions that each model attains the best performance. As can be seen, regarding all metrics, the Bayesian network outperformed other models on at least 62% of the total population. On the second place is the ANN, and it showed better performance on almost 30% of the total population regarding all metrics, except the MCE. In summary, this table showed that in more institutions and a more significant percentage of the dataset, the Bayesian network has superior capability in out-of-domain calibration compared to other employed models for agribusiness delinquency risk assessment.

4 DISCUSSION

Our analysis revealed that the most pronounced data drift was observed during changes in institutions, followed by state-level changes, albeit with a considerably smaller variation magnitude as detailed in Table 3. The fact that institutional changes encapsulate regional variances suggests that institutional dynamics, rather than merely regional factors, are the primary drivers of this shift. This aligns with the reality of the agribusiness, where different institutions often employ varying business rules and lack standardized practices. Such a finding underscores the importance of ensuring calibration across various distributions.

Table 7: Comparison between the Bayesian network, ANN, XGBoost, and LR models regarding the number of institutions and the percentage of the total population in which they obtained higher ECE, ACE, MCE, and Brier values.

Metric	BN		ANN		XGBoost		LR	
	#	%	#	%	#	%	#	%
ECE	**7**	**71.71**	2	28.29	0	0	0	0
ACE	**5**	**67.86**	4	32.14	0	0	0	0
MCE	**6**	**62.73**	2	5.29	0	0	1	31.98
Brier	**6**	**62.04**	3	37.96	0	0	0	0

To address these market realities, a robust modeling approach that can account for market inconsistencies is paramount. We have adopted a Bayesian network model, which allows the design of the model structure and the update with data of the model initially informed with priors. Thus, the Bayesian network can handle different institutional practices and regional variances. This approach ensures that our model remains robust and accurate even when subjected to the inconsistent realities of the market environment.

Given the data drift analysis, we implemented a Bayesian network specifically developed for agricultural loan assessments using our dataset. From the data presented in Table 1, we can draw several key conclusions regarding the performance of the Bayesian network compared to the baseline models across multiple metrics and institutions.

Superior Calibration: The Bayesian network outperforms the other models in terms of calibration, as indicated by the ECE and ACE metrics. With the Bayesian network yielding lower ECE and ACE values in 71.71% and 67.86% of the institutions respectively, it demonstrates better alignment between the predicted probabilities and observed frequencies on average, as well as across all predicted probabilities.

Resilience to Drastic Errors: The Bayesian network also stands out in terms of robustness to drastic prediction errors. As revealed by the MCE, the Bayesian network model exhibited the smallest maximum deviation in 62.73% of the institutions, surpassing the ANN, XGBoost, and LR models.

Overall Predictive Accuracy: The Bayesian network model also achieves superior results in terms of predictive accuracy, as denoted by the Brier Score. With the lowest mean squared difference between predicted probabilities and actual outcomes in 62.0% of the institutions, the Bayesian network model demonstrates greater overall prediction accuracy.

These findings reaffirm the robustness of the Bayesian network. Despite the diverse institutional practices and regional variances, the Bayesian network model consistently outperforms the ANN, XGBoost, and LR models in calibration, resilience to drastic prediction errors, and overall prediction accuracy across a majority of the institutions. Thus, it validates the effectiveness of our modeling approach in dealing with the market's inconsistent realities.

4.1 Future Research Direction

In the domain of credit risk assessment and banking services, the interpretability of model outputs is crucial, serving as a key driver of transparent and informed decision-making. As we chart our future research, our focus will be on expanding the interpretability of the Bayesian network model and assess how variations in feature importance impact model performance. Owing to the graphical structure of the Bayesian network, it has inherent potential for delivering clear interpretations. An essential part of our investigation will be examining the variability of feature importance across institutions to gain insights into unique institutional influences on credit risk.

We also plan to evaluate various network structure learning methods, aiming to optimize the capture of intricate dataset relationships. Furthermore, we will scrutinize the out-of-domain generalization capabilities of our Bayesian network model concerning data drift in terms of latent variables. Importantly, our research efforts will not only enhance our model's predictive accuracy and interpretability but will also contribute to private investment decision-making in agricultural finance. By identifying the factors that may impact credit risk performance, we can enable more informed and effective investment decisions and thereby contribute to the stability and growth of agricultural finance.

5 CONCLUSION

Our research presents compelling evidence for the superiority of the Bayesian network model in agricultural loan assessments. Notably, this model exhibits superior calibration, resilience to drastic prediction errors, and overall predictive accuracy, as compared to other benchmark models like the Artificial Neural Network (ANN), extreme Gradient Boosting (XGBoost), and Logistic Regression (LR). The ability of the Bayesian network to handle different institutional practices and regional variances, despite the lack of standardization and the presence of data drift, is noteworthy. The network outperforms its counterparts across most institutions, affirming its robustness and efficacy.

As we plan our future research, we aim further to research the interpretability of the Bayesian network model and how variations in feature importance impact model performance. Indeed, unique institutional influences on credit risk could offer valuable insights. Additionally, we plan to assess out-of-domain generalization capabilities of our model, while exploring various network structure learning methods. Our overarching goal is to enhance our model's predictive accuracy and interpretability. The resulting insights will contribute to improved decision-making in agricultural finance and bolster its stability and growth.

ACKNOWLEDGMENTS

This work has been supported and funded by Traive Inc. Moreover, the authors would like to thank Traive's agricultural credit risk assessment experts for their collaboration and feedback. Particularly, the authors thank Luis E.R. Lapo, Antonio H.S.O. Filho, and Rafael H. de Arruda. Also, the authors wish to thank the team that prepared the dataset for this research, including Abhay Sarda, Arthur G. Yuan, and Alan Martins.

REFERENCES

[1] Niklas Bussmann, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock. 2021. Explainable machine learning in credit risk management. *Computational Economics* 57, 1 (2021), 203–216. https://doi.org/10.1007/s10614-020-10042-0

[2] Henry Condori Alejo, Miguel Aceituno Rojo, and Guina Alzamora. 2021. Rural micro credit assessment using machine learning in a Peruvian microfinance institution. *Procedia Computer Science* 187 (01 2021), 408–413. https://doi.org/10.1016/j.procs.2021.04.117

[3] Pedro G. Fonseca and Hugo D. Lopes. 2017. Calibration of machine learning classifiers for probability of default modelling. arXiv:1710.08901

[4] Matthieu Garcin and Samuel Stéphan. 2021. Credit scoring using neural networks and SURE posterior probability calibration. arXiv:2107.07206 [q-fin.ST]

[5] Saman Ghaffarian, Mariska van der Voort, João Valente, Bedir Tekinerdogan, and Yann de Mey. 2022. Machine learning-based farm risk management: A systematic mapping review. *Computers and Electronics in Agriculture* 192 (2022), 106631. https://doi.org/10.1016/j.compag.2021.106631

[6] Charles Guan, Hendra Suryanto, Ashesh Mahidadia, Michael Bain, and Paul Compton. 2023. Responsible credit risk assessment with machine learning and knowledge acquisition. *Human-Centric Intelligent Systems* (2023), 1–12.

[7] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. arXiv:1706.04599

[8] Tanja Havemann, Christine Negra, and Fred Werneck. 2022. *Blended finance for agriculture: exploring the constraints and possibilities of combining fi nancial instruments for sustainable transitions.* Springer Nature Switzerland, Cham, 347–358. https://doi.org/10.1007/978-3-031-18560-1_23

[9] David Heckerman. 2008. *A tutorial on learning with Bayesian networks.* Springer Berlin Heidelberg, Berlin, Heidelberg, 33–82. https://doi.org/10.1007/978-3-540-85066-3_3

[10] Daphne Koller and Nir Friedman. 2009. *Probabilistic graphical models: Principles and techniques.* MIT press.

[11] Anil Kumar, Suneel Sharma, and Mehregan Mahdavi. 2021. Machine learning (ML) technologies for digital credit scoring in rural finance: A literature review. *Risks* 9, 11 (2021). https://doi.org/10.3390/risks9110192

[12] Aldo Levy and Riad Baha. 2021. Credit risk assessment: a comparison of the performances of the linear discriminant analysis and the logistic regression. *International Journal of Entrepreneurship and Small Business* 42, 1-2 (2021), 169–186. https://doi.org/10.1504/IJESB.2021.112265 arXiv:https://www.inderscienceonline.com/doi/pdf/10.1504/IJESB.2021.112265

[13] Hua Li, Yumeng Cao, Siwen Li, Jianbin Zhao, and Yutong Sun. 2020. XGBoost model and its application to personal credit evaluation. *IEEE Intelligent Systems* 35, 03 (may 2020), 52–61. https://doi.org/10.1109/MIS.2020.2972533

[14] ABID Lobna, Soukeina Zaghdene, Afif Masmoudi, and Sonia Zouari Ghorbel. 2017. Bayesian network modeling: A case study of credit scoring analysis of consumer loans default payment. *Asian Economic and Financial Review* 7, 9 (2017), 846.

[15] Khalil Masmoudi, Lobna Abid, and Afif Masmoudi. 2019. Credit risk modeling using Bayesian network with a latent variable. *Expert Systems with Applications* 127 (2019), 157–166. https://doi.org/10.1016/j.eswa.2019.03.014

[16] Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning* (Bonn, Germany) *(ICML '05)*. Association for Computing Machinery, New York, NY, USA, 625–632. https://doi.org/10.1145/1102351.1102430

[17] Tatjana Pavlenko and Oleksandr Chernyak. 2010. Credit risk modeling using bayesian networks. *International Journal of Intelligent Systems* 25, 4 (2010), 326–344.

[18] R. Perry, J. von Kügelgen*, and B. Schölkopf*. 2022. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, Vol. 35. Curran Associates, Inc., 10904–10917. https://proceedings.neurips.cc/paper_files/paper/2022/hash/46a126492ea6fb87410e55a58df2e189-Abstract-Conference.html *shared last author.

[19] Congjun Rao, Ming Liu, Mark Goh, and Jianghui Wen. 2020. 2-stage modified random forest model for credit risk assessment of P2P network lending to "Three Rurals" borrowers. *Applied Soft Computing* 95 (2020), 106570. https://doi.org/10.1016/j.asoc.2020.106570

[20] Si Shi, Rita Tse, Wuman Luo, Stefano D'Addona, and Giovanni Pau. 2022. Machine learning-driven credit risk: a systemic review. *Neural Computing and Applications* 34, 17 (2022), 14327–14339.

[21] Dirk Tasche. 2012. *The art of probability-of-default curve calibration.* Technical Report 1212.3716. https://ideas.repec.org/p/arx/papers/1212.3716.html

[22] Eric-Jan Wagenmakers, Michael Lee, Tom Lodewyckx, and Geoffrey J. Iverson. 2008. *Bayesian versus frequentist inference.* Springer New York, New York, NY, 181–207. https://doi.org/10.1007/978-0-387-09612-4_9

[23] Yuelin Wang, Yihan Zhang, Yan Lu, and Xinran Yu. 2020. A comparative assessment of credit risk model based on machine learning——a case study of bank loan data. *Procedia Computer Science* 174 (2020), 141–149.

[24] Ruixiang Zhang, Masanori Koyama, and Katsuhiko Ishiguro. 2020. Learning structured latent factors from dependent data: A generative model framework from information-theoretic perspective. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 11141–11152. https://proceedings.mlr.press/v119/zhang20m.html

[25] Na Zhao, Fengge Yao, and Fu-Sheng Tsai. 2022. Innovative mechanism of rural finance: Risk assessment methods and impact factors of agricultural loans based on personal emotion and artificial intelligence. *Journal of Environmental and Public Health* 2022, 1126489 (2022).

Received 24 July 2023; revised dd mm yyyy; accepted dd mm yyyy