UNIVERSIDAD NACIONAL DE SAN MARTIN

RCSI
https://revistas.unsm.edu.pe/index.php/rcsi
e-ISSN: 2709-992X

Original article
Artículo original
Jul-Dec, 2026

Prediction and calibration of payment default risk using machine learning in residential water service customers

Predicción y calibración del riesgo de impago mediante aprendizaje automático en usuarios residenciales de servicios de agua

Valles-Coral, Miguel Angel[1*]
Cuello-Sangama, Rusber[1]
Rengifo-Amasifen, Roger[2]
Vidaurre-Rojas, Pierre[1]
Prieto-Luna, Jaime Cesar[3]
Holgado-Apaza, Luis Alberto[3]
Injante, Richard[1]

[1]Universidad Nacional de San Martín, Tarapoto, Perú
[2]Universidad Nacional Autónoma de Alto Amazonas, Yurimaguas, Perú
[3]Universidad Nacional Amazónica de Madre de Dios, Puerto Maldonado, Perú

Received: 11 Dec. 2025 | **Accepted:** 30 Jun. 2026 | **Published:** 20 Jul. 2026
Corresponding author*: mavalles@unsm.edu.pe
How to cite this article: Valles-Coral, M.A., Cuello-Sangama, R., Rengifo-Amasifen, R., Vidaurre-Rojas, P., Prieto-Luna, J.C., Holgado-Apaza, L.A., Injante, R. (2026). Prediction and calibration of payment default risk using machine learning in residential water service customers. *Revista Científica de Sistemas e Informática*, *6*(2), e1265. https://doi.org/10.51252/rcsi.v6i2.1265

ABSTRACT

Water service payment delinquency limits the availability of resources required to sustain operations and hinders the prioritization of collection activities. This study aimed to evaluate machine learning models for estimating the payment default risk of residential customer bills at EMAPAB S.A. A total of 264,382 bills from 6,358 customers were analyzed using 51 predictors derived from historical billing and payment records. Random Forest, CatBoost, and a feedforward neural network were compared using Optuna optimization, stratified and grouped ten-fold cross-validation, and an independent temporal test. Random Forest achieved the best performance on the temporal test, with an accuracy of 0.9037, an F1-score of 0.6333, a ROC-AUC of 0.8681, and a PR-AUC of 0.6892. Its probabilities were subsequently calibrated using Platt Scaling and stratified into low-, medium-, and high-risk levels, accounting for 74.61%, 13.96%, and 11.44% of the bills, respectively. The results show that the proposed approach can identify bills with higher default risk and support the prioritization of the collection portfolio, although its operational impact requires prospective validation.

Keywords: billing; collection; payment behavior; stratification; temporal validation

RESUMEN

La morosidad en los servicios de agua limita la disponibilidad de recursos para sostener las operaciones y dificulta la priorización de las acciones de cobranza. El objetivo del estudio fue evaluar modelos de aprendizaje automático para estimar el riesgo de impago de recibos de usuarios residenciales de EMAPAB S.A. Se analizaron 264 382 recibos correspondientes a 6 358 usuarios, utilizando 51 predictores construidos a partir de registros históricos de facturación y pagos. Se compararon Random Forest, CatBoost y una red neuronal feedforward mediante optimización con Optuna, validación cruzada estratificada y agrupada de diez folds y una prueba temporal independiente. Random Forest obtuvo el mejor desempeño en el test temporal, con accuracy de 0.9037, F1-score de 0.6333, ROC-AUC de 0.8681 y PR-AUC de 0.6892. Posteriormente, sus probabilidades fueron calibradas mediante Platt Scaling y estratificadas en riesgo bajo, medio y alto, concentrando el 74.61 %, 13.96 % y 11.44 % de los recibos, respectivamente. Los resultados evidencian que el enfoque permite identificar recibos con mayor riesgo y puede apoyar la priorización de la cartera de cobranza, aunque su impacto operativo requiere validación prospectiva.

Palabras clave: cobranza; comportamiento de pago; estratificación; facturación; validación temporal

1. INTRODUCTION

The early identification of payment default risk has become increasingly important for organizations that manage recurring payments. In public utility companies, payment delays affect the availability of resources required to sustain operational activities and may increase the workload associated with collection processes. This issue is particularly relevant in water and sanitation services, where financial continuity depends largely on the timely recovery of billed amounts (Ayala esquivel & Cabrera Tapia, 2021; Montenegro-Chasquibol et al., 2024; Soudachanh et al., 2022).

The commercial systems used by these companies continuously generate information on consumption, billing, payments, and customer history. These records make it possible to analyze historical behavior and construct indicators associated with payment frequency, recency, and magnitude, creating opportunities to move from predominantly reactive management toward preventive support strategies. In this context, machine learning has been used to identify patterns in commercial data and support decision-making processes related to delinquency, segmentation, and payment behavior (Lara López et al., 2023; Lescano-Delgado, 2024; Yajure Ramírez, 2022).

The literature shows that different algorithms can take advantage of this type of information. Bashar et al. (2023) evaluated models from different algorithmic families to estimate the probability of payment among residential customers, while other studies have applied classification and segmentation techniques to analyze behaviors associated with billing, consumption, and compliance with payment obligations (García Hernández & Torres Moreno, 2023; Olmedo et al., 2023). Likewise, research conducted in contexts related to water services has shown that administrative data can be used to generate risk profiles and support the prioritization of commercial actions (Ramirez Villacorta, 2024).

However, a model's performance on historical data does not guarantee that it will behave similarly when applied to later periods. In longitudinal datasets, the distribution of the target variable and customer behavior patterns may change over time, making it necessary to assess generalization capability using a temporal data split. Similarly, when multiple records belong to the same customer, distributing them across training and validation sets may produce overly optimistic estimates if the dependence between observations is not properly controlled. These considerations become particularly relevant when a model is intended to support operational decision-making rather than serve solely as a classification exercise.

Another relevant aspect is the interpretation of the probabilities generated by predictive models. Adequate discrimination makes it possible to distinguish between higher- and lower-risk cases, but it does not guarantee that the estimated probabilities accurately represent the expected frequency of default. For this reason, probability calibration can complement predictive evaluation and facilitate the transformation of model outputs into risk levels that can be used to organize the collection portfolio. In this regard, the practical usefulness of a model depends not only on its ability to identify default cases, but also on the stability and interpretability of its estimates.

Although machine learning applications have been developed for payment behavior and customer management, there is less evidence in water utilities regarding approaches that simultaneously integrate historical features constructed without future information, customer-grouped validation, independent temporal evaluation, statistical comparison among algorithms, and probability calibration aimed at risk stratification. Such a combination may bring model evaluation closer to the conditions expected in eventual operational use.

In this context, the objective of this study was to evaluate machine learning models for estimating the payment default risk of bills issued to residential customers of the Municipal Drinking Water and Sewerage Company of Bagua S.A. Random Forest, CatBoost, and a feedforward neural network were compared using stratified cross-validation grouped by customer, followed by an independent temporal evaluation. Finally,

the probabilities generated by the selected model were calibrated and organized into risk levels to support the prioritization of the collection portfolio.

2. MATERIALS AND METHODS

2.1 Study Design and Data Source

The study followed an applied quantitative approach and was conducted through a retrospective analysis of administrative records from the Municipal Drinking Water and Sewerage Company of Bagua S.A. (EMAPAB S.A.), located in Bagua, Amazonas, Peru. The unit of analysis was each billing record corresponding to residential customers that met the established processing criteria. The information was obtained from the institutional commercial system, following authorization from the company and coordination with the Commercial and Systems departments.

The data extracted from the institutional system and the main sources used in the study are summarized in Figure 1.

Figure 1. Data sources and information components used in the analysis

Following inspection, cabfacturacion, cabpagos, and detpagos were used as the main data sources because they contained the information required on billing, payments, and transactions associated with each bill. To preserve confidentiality, names, addresses, and other directly identifying information were excluded. Record integration was performed using technical system identifiers, including company code, branch code, registration number, and billing number.

2.2 Data Preparation and Definition of the Target Variable

Billing and payment records underwent a cleaning and data-type standardization process prior to integration, as data quality and consistency are relevant aspects of predictive modeling (Li et al., 2018). Quantitative variables were converted to numeric format, and dates were standardized to enable chronological ordering of records by customer. Inconsistent periods and dates, negative amounts, voided payments, and duplicate records were also reviewed. Extreme values were inspected but retained because they could correspond to actual consumption, billing, or payment transactions (Abuzaid & Alkronz, 2024).

The target variable was defined at the bill level by linking billing records with payment details using a composite key consisting of company code, branch code, registration number, and billing number. Bills were considered eligible when they had complete identifiers, an original amount greater than zero, valid issue and due dates, and a complete follow-up period extending to 30 days after the due date.

For each bill, all valid payments recorded up to the cutoff date were accumulated, and the outstanding balance was calculated. impago_real = 0 was assigned when the remaining balance was less than or equal to PEN 0.01, whereas impago_real = 1 was assigned when the bill retained an outstanding balance greater than this amount. Records without a complete observation period were excluded from the supervised dataset.

2.3 Feature Engineering and Data Leakage Control

Billing and payment records were organized by customer and period to construct variables representing prior behavior. Billing information included billed amounts, consumption, and number of bills, while payment information was consolidated according to the amount paid, number of transactions, and presence of recorded payments. Based on these data, features were generated to represent periods without payment, payment-to-billing ratios, time elapsed since the most recent payment, as well as sums, averages, and standard deviations calculated over historical windows of 3, 6, and 12 periods.

To ensure that each prediction used only information available before the bill being evaluated, historical variables were shifted by one period using shift (1) prior to calculating the temporal windows. These features were then integrated with eligible bills using the registration number and billing period. Predictor fields associated with subsequent payments, outstanding balance, final bill status, and any variable used to construct the default label were excluded. In addition, indicator variables were incorporated to distinguish the absence of historical information from cases in which the observed value was effectively zero.

2.4 Dataset Composition and Temporal Partitioning

The dataset used for modeling consisted of 264,382 bills corresponding to 6,358 residential customers (Table 1). The integrated dataset contained 60 variables, of which 51 were used as predictors: 49 numerical and two categorical variables (codzona and codtipousuario). The remaining columns corresponded to identifiers, temporal control variables, and the target variable. Predictors were organized according to their availability and construction method into three groups: current, temporal, and availability variables; historical availability, lag, and recency variables; and historical features calculated using moving windows.

Table 1. Composition of the predictor variables used for modeling

Predictor group	Number	Description
Current, temporal, and availability variables	12	Features available at the time the bill was issued, temporal components, and indicators reflecting the availability of historical information.
Historical availability, lag, and recency variables	12	Features associated with customer history, values from the previous period, and time elapsed since previous payments.
Historical variables calculated using moving windows	27	Billing, consumption, and payment statistics constructed from the previous 3, 6, and 12 periods.
Total	**51**	**49 numerical and 2 categorical variables used as input to the predictive models.**

The dataset was divided chronologically to preserve the temporal sequence of the records. Bills corresponding to periods up to September 2022 were assigned to model training; records from October to December 2022 were reserved for hyperparameter optimization; and those corresponding to January–March 2023 formed the temporal test set. The resulting distribution is presented in Table 2.

Table 2. Composition of the temporal dataset partitions

Partition	Period	Records	Customers	Default
Training	Up to September 2022	230.126	6.262	31.84 %

Hyperparameter optimization	October–December 2022	17.075	5.718	12.30 %
Temporal test	January–March 2023	17.181	5.795	13.18 %

The difference in default proportions across partitions indicated temporal variation in the distribution of the target variable. For this reason, the January–March 2023 dataset was kept separate from the optimization and internal validation stages in order to subsequently evaluate model performance on records from later periods.

2.5 Predictive Models and Hyperparameter Optimization

Three supervised classification models were evaluated: Random Forest (Breiman, 2001), CatBoost (Prokhorenkova et al., 2018) and a feedforward neural network (FFNN) (Hornik et al., 1989). Preprocessing was adapted to each algorithm and executed within its corresponding pipeline. For Random Forest, missing numerical values were imputed using the median and categorical variables were encoded, without applying standardization. CatBoost directly handled categorical variables and incorporated early stopping during training, whereas in the FFNN numerical variables were imputed and standardized using StandardScaler.

Hyperparameters were optimized with Optuna through 30 trials per model, using PR-AUC as the objective function. The dataset allocated to this stage was divided into two customer-grouped subsets: 13,667 bills for fitting and 3,408 for evaluating the configurations, preventing records from the same customer from appearing simultaneously in both groups. For each algorithm, the configuration achieving the highest PR-AUC was retained. The main selected hyperparameters are presented in Table 3.

Table 3. Selected hyperparameters of the evaluated models

Model	Selected configuration
Random Forest	800 trees; maximum depth of 16; minimum of 14 samples required to split a node and 7 samples per leaf; entropy criterion; max_features = sqrt; max_samples = 0.85.
CatBoost	516 iterations; learning rate of 0.0112; depth of 9; L2 regularization of 17.1576; Bernoulli bootstrap; subsample of 0.6173.
FFNN	Three hidden layers with 224, 96, and 192 neurons; ReLU activation; learning rate of 0.00122; batch size of 256; sigmoid output.

2.6 Validation and Predictive Performance Evaluation

The selected Random Forest, CatBoost, and FFNN configurations were evaluated using stratified, grouped ten-fold cross-validation, following an evaluation scheme aimed at obtaining more robust estimates of predictive performance (Hastie et al., 2009). The customer registration number was used as the grouping variable so that all bills associated with the same customer remained within a single partition during each iteration. The three models used the same folds, and in each cycle, nine partitions were used for training and one for validation. Predictions obtained across the ten folds were combined to generate out-of-fold (OOF) predictions for each model.

Performance was evaluated using accuracy, balanced accuracy, precision, recall, F1-score for the default class, specificity, ROC-AUC, PR-AUC, log-loss, Brier score, and the Matthews correlation coefficient, which assess different aspects of classifier performance (Tharwat, 2021). A common threshold of 0.50 was used for the initial comparison; subsequently, model-specific thresholds were determined from the OOF probabilities by maximizing the F1-score for the default class. Fold-level metrics, estimated probabilities, and resulting predictions were retained for the subsequent statistical analysis.

2.7 Statistical Analysis and Temporal Evaluation

The statistical comparison of the models was based on the F1-scores obtained across the same ten validation folds. Given the paired nature of the observations, the Friedman test was applied to identify overall differences among Random Forest, CatBoost, and FFNN, followed by pairwise comparisons using the Wilcoxon signed-rank test and Holm adjustment to control significance across multiple comparisons (García et al., 2010). Kendall's W coefficient was also calculated as a measure of effect size.

The final model was selected using the highest mean F1-score obtained during grouped cross-validation as the primary criterion, while the remaining metrics were considered complementary indicators of performance. The models were subsequently evaluated on the independent temporal test set corresponding to January–March 2023, without any adjustment using these records. This separation made it possible to assess predictive behavior on temporally subsequent observations while accounting for the temporal structure and dependence among records belonging to the same customer (Roberts et al., 2017). For model comparisons during this period, a customer-grouped bootstrap procedure was applied, keeping all bills associated with the same registration number together.

2.8 Probability Calibration and Risk Stratification

The OOF probabilities generated by the selected model were used to evaluate three calibration alternatives: uncalibrated probabilities, Platt Scaling, and isotonic regression. The quality of each alternative was compared using the Brier score, log-loss, and expected calibration error (ECE), which were used to assess the agreement between estimated probabilities and observed outcomes (Mervin et al., 2020; Odesola et al., 2025). The selected calibrator was fitted exclusively using the OOF predictions and was subsequently applied, without further adjustment, to the probabilities generated for the temporal test set, thereby preserving the independence of the final evaluation.

Based on the calibrated OOF probabilities, three risk levels were established considering the relationship between precision and recall: low risk for values below 0.200120, medium risk for probabilities from 0.200120 to values below 0.429122, and high risk for values equal to or greater than 0.429122.

Independently, a binary threshold of 0.388105 was determined by maximizing the F1-score for the default class. The risk levels were used to stratify the bills, whereas the binary threshold was used to evaluate classification performance on the temporal test set.

3. RESULTS AND DISCUSSION

3.1 Performance of the Models in Cross-Validation

Evaluation using stratified and grouped ten-fold cross-validation showed similar performance among the three models, although differences were observed in the balance between precision and default detection. Random Forest achieved the highest mean F1-score (0.7691), precision (0.8512), and PR-AUC (0.8831), whereas the FFNN achieved the highest recall (0.7156). CatBoost obtained results close to those of Random Forest across the main performance metrics (Table 4).

Table 4. Mean performance of the models in grouped ten-fold cross-validation

Model	Accuracy	Balanced accuracy	Precision	Recall	F1-Score	PR-AUC
Random Forest	0.8660	0.8222	**0.8512**	0.7016	**0.7691**	**0.8831**
CatBoost	0.8629	0.8219	0.8351	0.7093	0.7670	0.8777
FFNN	0.8568	0.8191	0.8124	**0.7156**	0.7607	0.8666

The observed behavior revealed a trade-off between precision and sensitivity. The FFNN identified a larger proportion of defaulted bills but showed lower precision and F1-score, whereas Random Forest maintained a more favorable balance between both metrics. CatBoost showed intermediate performance and remained very close to Random Forest, with a difference of only 0.0021 in mean F1-score.

The discriminatory ability derived from the out-of-fold predictions was also high for all three algorithms, with ROC-AUC values of 0.9282 for Random Forest, 0.9247 for CatBoost, and 0.9175 for the FFNN. Overall, these results indicated that tree-based models retained a slight advantage during internal validation, although the magnitude and statistical significance of these differences required the paired statistical analysis described below.

3.2 Statistical Comparison of the Models

The distribution of F1-scores across the ten folds showed very similar behavior between Random Forest and CatBoost, whereas the FFNN remained below both models across the evaluated partitions. This trend was consistent with the previously observed averages and showed that the advantage of the tree-based models was not concentrated in a single fold but persisted throughout the validation process (Figure 2).

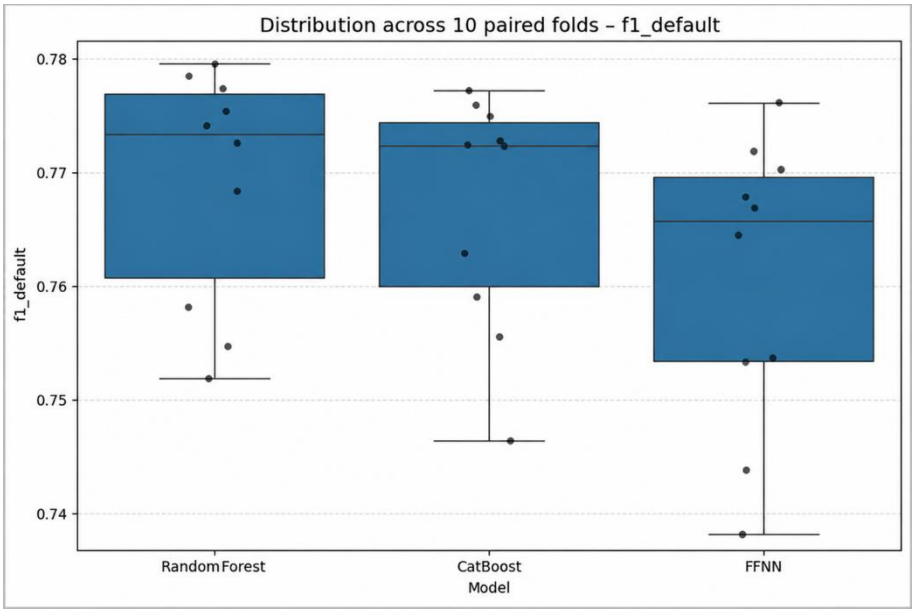

Figure 2. Distribution of model F1-scores across the ten cross-validation folds

The overall comparison confirmed that the differences observed among the three models were statistically significant. In addition, Kendall's W indicated that the separation in their performance was consistent across the analyzed partitions (Table 5).

Table 5. Overall comparison of F1-score using the Friedman test

Friedman statistic	Degrees of freedom	p-value	Kendall's W	Result
15.800	2	< 0.001	0.790	Significant difference

However, pairwise comparisons showed that the overall difference was mainly associated with the performance of the FFNN. Random Forest and CatBoost did not differ significantly from each other, indicating comparable internal performance. In contrast, the differences between Random Forest and the FFNN, as well as between CatBoost and the FFNN, were significant, confirming a consistent advantage of the tree-based models during cross-validation (Table 6).

Table 6. Pairwise F1-score comparisons using the Wilcoxon signed-rank test with Holm correction

Comparison	Mean difference	Adjusted p-value	Significant difference
Random Forest – CatBoost	0.0021	0.0840	No
Random Forest – FFNN	0.0085	0.0059	Yes
CatBoost – FFNN	0.0064	0.0059	Yes

Overall, the results did not establish a statistically significant superiority of Random Forest over CatBoost during internal validation, although Random Forest maintained the highest mean F1-score. This similarity made it necessary to examine model behavior on temporally subsequent records before assessing their generalization capability.

3.3 Performance in the Temporal Evaluation

Evaluation on records from January to March 2023 showed a decline in the performance of all three models compared with internal validation, although the magnitude of the decline varied (Table 7). Random Forest maintained the most favorable balance between default detection and precision, whereas CatBoost and, particularly, the FFNN increased recall at the cost of a larger number of false alerts. This behavior was particularly relevant considering that the default rate in the temporal dataset was lower than that observed during training.

Table 7. Performance of the models on the temporal test set

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC	PR-AUC
Random Forest	0.9037	0.6357	0.6309	0.6333	0.8681	0.6892
CatBoost	0.8651	0.4915	0.6733	0.5682	0.8514	0.6570
FFNN	0.8153	0.3943	0.7475	0.5162	0.8625	0.6370

The comparison between out-of-fold validation and the temporal test set showed that the reduction in F1-score was smaller for Random Forest (Figure 3). In contrast, CatBoost and the FFNN exhibited a more pronounced decline, suggesting greater sensitivity to changes in records from later periods. Although the FFNN identified a higher proportion of defaults, its lower precision reduced its usefulness for potential collection prioritization because it generated a considerably larger number of cases incorrectly classified as defaults.

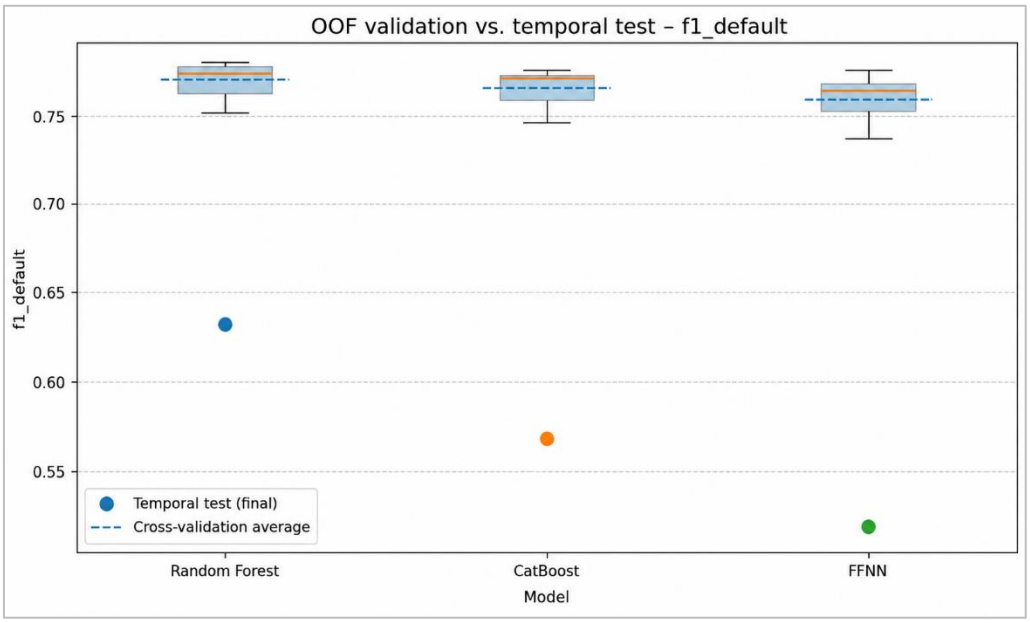

Figure 3. Comparison of F1-score obtained from out-of-fold validation and the temporal test set

Customer-grouped bootstrap comparisons showed significant differences among the models during the temporal period, with adjusted p-values of 0.003. Under these conditions, Random Forest outperformed CatBoost and the FFNN in terms of F1-score, whereas CatBoost performed better than the FFNN. Overall, the temporal evaluation supported the previous selection of Random Forest and showed greater stability in its performance when applied to later records.

3.4 Importance of Predictor Variables

The feature importance analysis of the Random Forest model showed a predominance of variables related to historical payment behavior. Among the predictors with the greatest contribution were the number of periods without payment in recent windows, payment-to-billing ratios, and amounts paid in previous periods (Table 8). Overall, this pattern indicates that recent payment behavior provided greater predictive value than isolated administrative or billing characteristics.

Table 8. Variables with the highest importance in the Random Forest model

Rank	Variable	Importance
1	periodos_sin_pago_6	0.0790
2	periodos_sin_pago_3	0.0649
3	ratio_pago_facturacion_3_periodos	0.0649
4	pago_prom_3_periodos	0.0430
5	n_pagos_sum_12_periodos	0.0427
6	periodos_sin_pago_12	0.0415
7	ratio_pago_facturacion_6_periodos	0.0381
8	dias_hasta_vencimiento	0.0371
9	n_pagos_sum_3_periodos	0.0369
10	pago_sum_3_periodos	0.0344

The simultaneous presence of variables calculated over 3-, 6-, and 12-period windows suggests that the model captured both recent patterns and longer-term historical behavior. However, these values should be interpreted only as relative contributions within the model and not as evidence of causal relationships with payment default.

3.5 Probability Calibration and Risk Stratification

The comparison of calibration alternatives showed differences in the quality of the probabilities generated by Random Forest. Platt Scaling achieved the lowest Brier score and log-loss values, whereas isotonic regression achieved the lowest expected calibration error (ECE). According to the predefined criterion, Platt Scaling was selected as the final calibrator and subsequently applied, without further adjustment, to the independent temporal dataset (Table 9).

Table 9. Performance of calibration methods on Random Forest OOF predictions

Method	Brier score	Log-loss	ECE
Uncalibrated	0.095695	0.308345	0.021557
Platt Scaling	0.094946	0.305202	0.002539
Isotonic regression	0.094966	0.305249	0.000970

The retrospective application of the calibrated model maintained a high ability to identify non-defaulted bills and a moderate balance between precision and detection of the positive class. Using the binary threshold of 0.388105, previously determined from the OOF predictions, an F1-score of 0.6319 and a specificity of 0.9468 were obtained, without altering the discriminatory ability expressed by ROC-AUC and PR-AUC (Table 10). These results correspond to the application of the model and calibrator to 17,181 bills in the temporal dataset without any further adjustment.

Table 10. Retrospective performance of the calibrated Random Forest model on the temporal dataset

Accuracy	Balanced accuracy	Precision	Recall	F1-score	Specificity	ROC-AUC	PR-AUC
0.9042	0.7853	0.6402	0.6238	0.6319	0.9468	0.8681	0.6892

Probability quality also remained adequate during the temporal period, with a Brier score of 0.0704, log-loss of 0.2585, and ECE of 0.0471. These metrics made it possible to examine the model's probabilistic behavior beyond its classification performance.

Subsequent stratification showed that 12,818 bills (74.61%) were classified as low risk, 2,398 (13.96%) as medium risk, and 1,965 (11.44%) as high risk (Figure 4). Together, the medium- and high-risk groups comprised 4,363 bills, equivalent to 25.40% of the temporal dataset.

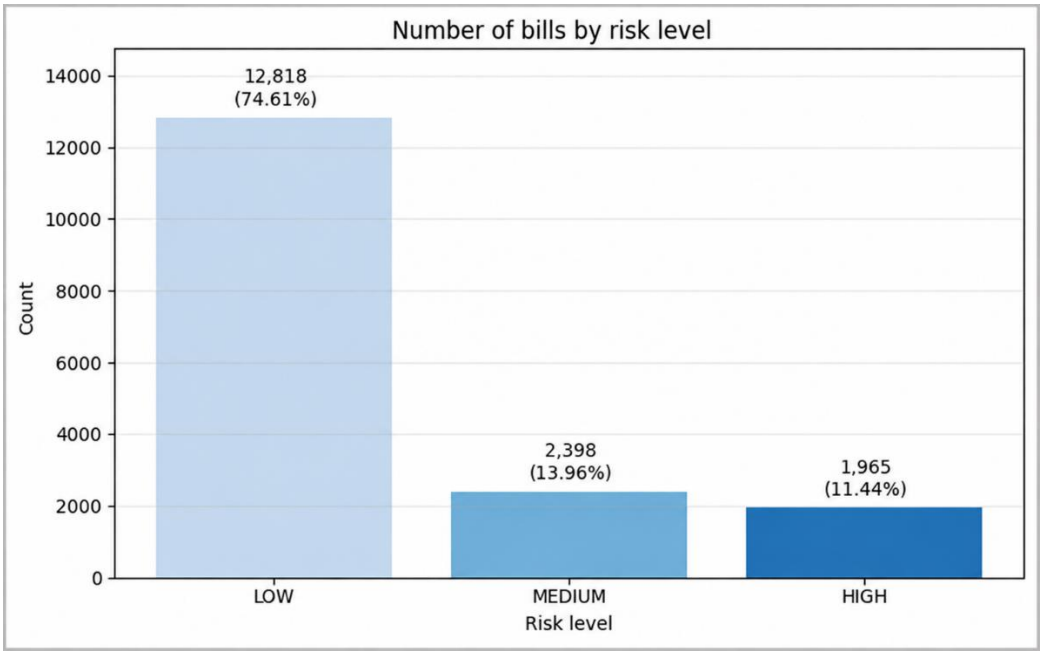

Figure 4. Number of bills according to payment default risk level

This distribution shows that calibrated probabilities can be used to rank the collection portfolio and identify a subset of bills with potentially higher priority for follow-up. However, the resulting risk levels represent a decision-support tool and do not, by themselves, demonstrate an improvement in debt recovery, as their operational usefulness must be confirmed through prospective application within the collection process.

3.6 Discusión

The results showed that Random Forest provided the most favorable balance for estimating payment default risk, particularly when applied to records from a later period. This finding is consistent with Bashar et al. (2023), who reported competitive performance of Random Forest in predicting the propensity to pay among residential energy customers. It also aligns with García Hernández and Torres Moreno (2023) regarding the usefulness of machine learning for anticipating nonpayment behavior based on historical records. However, during internal validation, Random Forest and CatBoost achieved similar performance without statistically significant differences, indicating that the advantage of Random Forest was not absolute but was mainly reflected in a better balance between precision and default detection and in greater stability when applied to temporally subsequent data.

The decline in F1-score observed for all three algorithms during temporal evaluation showed that performance estimated through internal validation may decrease when the application period changes. This effect was smaller for Random Forest, whereas CatBoost and, particularly, the FFNN increased default detection at the cost of generating a larger number of false alerts. From an operational perspective, this behavior is relevant because higher sensitivity does not necessarily represent a better alternative when it leads collection efforts to be directed toward numerous bills that are ultimately paid. Therefore, the independent temporal evaluation provided a more realistic assessment of generalization under conditions closer to the model's potential operational use.

Feature importance showed that risk was mainly associated with previous payment behavior, particularly periods without payment, payment-to-billing ratios, and amounts paid within historical windows. This finding reinforces the usefulness of commercial records for characterizing customer behavior patterns. Studies by Olmedo et al. (2023), Lara López et al. (2023), and Yajure Ramírez (2022) have shown that historical consumption and billing information can identify differentiated behavioral patterns in public utilities, although their objectives were not specifically focused on predicting payment default. In the present study, the use of 3-, 6-, and 12-period windows made it possible to represent both recent behavior and more persistent historical patterns.

Finally, calibration through Platt Scaling transformed Random Forest outputs into probabilities that could be organized into risk levels. The concentration of 25.40% of bills in the medium- and high-risk strata suggests that the model could help identify a priority group for follow-up rather than applying uniform actions across the entire portfolio. This potential usefulness should be interpreted as decision support rather than evidence of an actual reduction in delinquency, because the study was retrospective and did not evaluate real collection interventions. Accordingly, a subsequent stage should prospectively validate the model and determine whether risk-based prioritization produces measurable improvements in payment recovery.

CONCLUSIONS

The study established a predictive approach for estimating payment default risk based on the historical billing and payment behavior of residential customers of EMAPAB S.A. Among the evaluated algorithms, Random Forest showed the most suitable conditions for application by maintaining consistent performance on temporally subsequent data. Incorporating payment history across different time horizons made it possible to capture relevant patterns in customer behavior, while probability calibration facilitated the transformation of model estimates into interpretable risk levels. This stratification provides an alternative for directing follow-up efforts toward bills with a higher probability of default and supporting data-driven collection management. However, its operational usefulness should be confirmed through prospective implementation to determine its actual effect on payment recovery and to update the model in response to potential changes in customer behavior.

FINANCING

The authors received no financial support or sponsorship for conducting this study or preparing this article.

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest related to the subject matter of this study.

AUTHORSHIP CONTRIBUTION

Conceptualization: Valles-Coral, M. A.; Cuello-Sangama, R. Methodology: Valles-Coral, M. A.; Cuello-Sangama, R. Investigation: Cuello-Sangama, R.; Prieto-Luna, J. C. Supervision: Valles-Coral, M. A.; Vidaurre-Rojas, P. Validation: Valles-Coral, M. A.; Holgado-Apaza, L. A. Data Curation: Rengifo-Amasifen, R.; Prieto-Luna, J. C. Formal Analysis: Rengifo-Amasifen, R.; Holgado-Apaza, L. A. Writing – Review and Editing: Vidaurre-Rojas, P.

REFERENCES

Abuzaid, A., & Alkronz, E. (2024). A comparative study on univariate outlier winsorization methods in data science context. *Italian Journal of Applied Statistics*, *36*(1). https://doi.org/10.26398/IJAS.0036-004

Ayala esquivel, B. D., & Cabrera Tapia, C. F. (2021). La importancia de la economía del agua. *RD-ICUAP*, 78–91. https://doi.org/10.32399/icuap.rdic.2448-5829.2021.21.630

Bashar, M. A., Nayak, R., Astin-Walmsley, K., & Heath, K. (2023). Machine learning for predicting propensity-to-pay energy bills. *Intelligent Systems with Applications*, *17*, 200176. https://doi.org/10.1016/j.iswa.2023.200176

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

García Hernández, J. M., & Torres Moreno, W. N. (2023). *Predicción de riesgo de impago en institución financiera usando modelos de Machine Learning* [Universidad Tecnológica Centroamericana UNITEC]. https://repositorio.unitec.edu/xmlui/handle/123456789/12906

García, S., Fernández, A., Luengo, J., & Herrera, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences*, *180*(10), 2044–2064. https://doi.org/10.1016/j.ins.2009.12.010

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Lara López, F., Manríquez García, N., & Quintero Rodríguez, J. O. (2023). Comportamiento de la demanda del consumo de agua potable por zonas en Mazatlán, Sinaloa. *INTER DISCIPLINA*, *11*(31), 317–337. https://doi.org/10.22201/ceiich.24485705e.2023.31.86085

Lescano-Delgado, M. (2024). Avances en el uso de inteligencia artificial para la mejora del control y la detección de fraudes en organizaciones. *Revista Científica de Sistemas e Informática*, *4*(2), e671. https://doi.org/10.51252/rcsi.v4i2.671

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2018). Feature Selection. *ACM Computing Surveys*, *50*(6), 1–45. https://doi.org/10.1145/3136625

Mervin, L., H., Afzal, A. M., Engkvist, O., & Bender, A. (2020). Comparison of Scaling Methods to Obtain Calibrated Probabilities of Activity for Protein–Ligand Predictions. *Journal of Chemical Information and Modeling*, *60*(10), 4546–4559. https://doi.org/10.1021/acs.jcim.0c00476

Montenegro-Chasquibol, B., Labajos-Portocarrero, H., Terrones-Suarez, O., Minga-Sarmiento, R., & Fasanando-Garcia, S. (2024). Gestión de cobranzas y su influencia en la liquidez de una empresa inmobiliaria. *Revista Amazónica de Ciencias Económicas*, *3*(2), e741. https://doi.org/10.51252/race.v3i2.741

Odesola, P. A., Adegoke, A. A., & Babalola, I. (2025). *Model uncertainty quantification: A post hoc calibration approach for heart disease prediction*. https://doi.org/10.1101/2025.09.28.25336834

Olmedo, A. T., Castro, J. R., Reátegui, R., & Castillo, T. (2023). Aplicación de algoritmos de Machine Learning para la segmentación del consumo de agua potable. Caso de estudio en Catamayo, Ecuador. *2023 18th Iberian Conference on Information Systems and Technologies (CISTI)*, 1–6. https://doi.org/10.23919/CISTI58278.2023.10211900

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems 31*.

Ramirez Villacorta, J. M. (2024). *Modelo predictivo basado en redes neuronales artificiales para pronosticar el consumo de agua potable en la ciudad de Iquitos* [Universidad Nacional Federico Villarreal]. https://hdl.handle.net/20.500.13084/8532

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929. https://doi.org/10.1111/ecog.02881

Soudachanh, S., Langergraber, G., & Salhofer, S. (2022). Sanitation planning for resettlement sites in Laos. *Journal of Water, Sanitation and Hygiene for Development*, *12*(3), 248–257. https://doi.org/10.2166/washdev.2022.178

Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003

Yajure Ramírez, C. A. (2022). Uso de algoritmos de Machine Learning para analizar los datos de energía eléctrica facturada en la Ciudad de Buenos Aires durante el período 2010–2021. *Ciencia, Ingenierías y Aplicaciones*, *5*(2), 7–37. https://doi.org/10.22206/cyap.2022.v5i2.pp7-37