UDC 004.8:336.71
DOI **https://doi.org/10.52171/herald.373**

Improving the Reliability of Bank Customer Churn Prediction via Calibration and Uncertainty Quantification

M.E. Rahimov

Azerbaijan Technical University (Baku, Azerbaijan)

For correspondence:
Musa Rahimov / e-mail: musa.rahimov@aztu.edu.az

Abstract

This study looks into how machine learning models for customer churn prediction in the banking industry use uncertainty quantification and calibration analysis. With an emphasis on predictive accuracy, probabilistic calibration, and profitability, the study contrasts three sophisticated models: Random Forest (Deep Ensemble), XGBoost, and a Neural Network with Monte Carlo Dropout. The assessment took into account both traditional and reliability-focused criteria, such as AUC, F1-score, Brier score, projected Calibration Error (ECE), and projected profit. The Random Forest model produced the most profit (118,000 AZN) with an AUC of 0.842 and an F1-score of 0.777. XGBoost displayed a moderate calibration variation (ECE = 0.088) but somewhat increased the F1-score to 0.794. On the other hand, while having a lower F1-score (0.631), the NN + MCDO model achieved the greatest AUC value (0.863) and the best calibration consistency (ECE = 0.026). The results indicate that uncertainty-aware deep learning improves probabilistic reliability and interpretability, whereas ensemble-based models are more effective for profit optimization in banking decision systems.

Keywords: Banking, Machine Learning, Deep Learning, Random Forest, XGBoost, Monte Carlo Dropout, Calibration

Submitted *18 January 2026*
Published online *22 January 2026*

For citation:
M.E. Rahimov
[Improving the Reliability of Bank Customer Churn Prediction via Calibration and Uncertainty Quantification]
Herald of the Azerbaijan Engineering Academy, 2026, *ONLINE*
ISSN 2789-8245, CC BY-NC 4.0 https://creativecommons.org/licenses/by-nc/4.0/

Kalibrləmə və qeyri-müəyyənliyin qiymətləndirilməsi vasitəsilə bank müştərisi itkisinin proqnozlaşdırılmasının etibarlılığının artırılması

M.E. Rəhimov
Azərbaycan Texniki Universiteti (Bakı, Azərbaycan)

Xülasə

Bu tədqiqat bank sektorunda müştəri itkisini (churn) proqnozlaşdırmaq üçün istifadə olunan maşın öyrənməsi modellərində qeyri-müəyyənliyin qiymətləndirilməsi və kalibrasiya analizinin rolunu araşdırır. İş çərçivəsində proqnoz dəqiqliyi, ehtimal çıxışlarının etibarlılığı və iqtisadi səmərəlilik baxımından üç qabaqcıl model – Random Forest (Deep Ensemble), XGBoost və Monte Karlo Dropout ilə təkmilləşdirilmiş Neyron Şəbəkəsi müqayisəli şəkildə təhlil edilmişdir. Qiymətləndirmə prosesində həm ənənəvi performans göstəriciləri, həm də etibarlılıq yönümlü metriklər, o cümlədən AUC, F1-ölçü, Brier göstəricisi, Gözlənilən Kalibrasiya Xətası (ECE) və gözlənilən mənfəət nəzərə alınmışdır. Nəticələrə əsasən, Random Forest modeli 0.842 AUC və 0.777 F1-ölçü ilə ən yüksək iqtisadi nəticəni (118,000 AZN) təmin etmişdir. XGBoost modeli daha yüksək F1-ölçü (0.794) göstərsə də, orta səviyyəli kalibrasiya sapması (ECE = 0.088) nümayiş etdirmişdir. Digər tərəfdən, NN + MCDO modeli nisbətən aşağı F1-ölçüyə (0.631) malik olsa da, ən yüksək AUC göstəricisini (0.863) və ən yaxşı kalibrasiya uyğunluğunu (ECE = 0.026) əldə etmişdir. Nəticələr göstərir ki, qeyri-müəyyənliyi nəzərə alan dərin öyrənmə yanaşması ehtimal əsaslı proqnozların etibarlılığını və interpretasiya imkanlarını artırır, ansambl əsaslı modellər isə bank qərarvermə sistemlərində mənfəətin optimallaşdırılması baxımından daha effektivdir.

Açar sözlər: bankçılıq, maşın öyrənməsi, dərin öyrənmə, Random Forest, XGBoost, Monte Carlo Dropout, kalibrasiya

Повышение надежности прогнозирования оттока клиентов в банковской сфере на основе калибровки и количественной оценки неопределенности

М.Э. Рахимов
Азербайджанский Технический Университет (Баку, Азербайджан)

Аннотация

В данном исследовании анализируется роль калибровки и количественной оценки неопределенности в моделях машинного обучения, используемых для прогнозирования оттока банковских клиентов (churn). Точность и надежность прогнозирования решений клиентов имеют критическое значение для стратегического планирования и оптимизации прибыли в банковском секторе. В рамках исследования были применены три модели машинного обучения: Random Forest (глубокий ансамбль), XGBoost и нейронная сеть с использованием Monte Carlo Dropout. Оценка моделей проводилась на основе как традиционных метрик производительности, так и показателей калибровки и неопределенности, включая AUC, F1-меру, показатель Брайера, ожидаемую ошибку калибровки (ECE) и прогнозируемую прибыль. Согласно полученным результатам, модель Random Forest продемонстрировала наивысшую прибыль (118 000 AZN) при сбалансированных значениях AUC (0,842) и F1-меры (0,777). Модель XGBoost показала наивысшее значение F1-меры (0,794), однако характеризовалась умеренной ошибкой калибровки (ECE = 0,088). В свою очередь, модель NN + MCDO обеспечила наилучшее соответствие вероятностных прогнозов реальным частотам, продемонстрировав минимальную ошибку калибровки (ECE = 0,026) и наивысшее значение AUC (0,863), несмотря на более низкое значение F1-меры (0,631). Результаты показывают, что учет неопределенности в моделях глубокого обучения повышает надежность и интерпретируемость вероятностных прогнозов, тогда как ансамблевые модели более эффективны с точки зрения оптимизации прибыли в системах банковского принятия решений.

Ключевые слова: банковский сектор, машинное обучение, глубокое обучение, Random Forest, XGBoost, Monte Carlo Dropout, калибровка

Introduction

Customer churn prediction is now a crucial part of data-driven customer relationship management in today's fiercely competitive financial landscape. Banks can create proactive retention strategies and reduce any financial losses by identifying clients who are likely to quit. However, the efficacy of conventional statistical methods for churn prediction is limited due to the introduction of complex and nonlinear interactions in banking data brought about by growing digitization and changing consumer behavior patterns [1].

In order to better capture such complex patterns and enhance prediction performance in customer churn analysis, machine learning (ML) and deep learning (DL) approaches have been widely utilized [2]. Similar benefits of machine learning methods have also been documented in general classification issues in other application domains, demonstrating how well they can extract significant patterns from complicated datasets [3]. Because of their resilience, capacity to understand feature interactions, and comparatively consistent generalization performance, ensemble-based models like Random Forest (RF) and XGBoost have proven to be highly predictive in banking applications [1, 4].

Recent research has revealed that many high-performance machine learning models have a tendency to generate poorly calibrated probability estimates, frequently displaying overconfident predictions that may lower the dependability of decision-making in financial systems, despite their strong discriminative capability [5]. Growing interest in calibration analysis and uncertainty quantification as supplementary evaluation dimensions beyond traditional accuracy-oriented measures has been spurred by this constraint.

In order to increase the reliability of probabilistic predictions in practical applications, proper calibration seeks to guarantee consistency between expected probability and observed outcomes. In this regard, deep neural networks enhanced with Monte Carlo Dropout (MCDO) have drawn interest as a useful approximation for Bayesian inference that makes it possible to estimate the epistemic uncertainty related to model parameters [5]. In high-risk financial situations, where accurate prediction is just as important as accurate confidence estimation, this kind of uncertainty awareness is especially pertinent.

Few studies have combined discriminative performance, calibration quality, and uncertainty estimates inside a single evaluation framework for banking churn prediction, despite recent research showing significant advancements in churn prediction utilizing sophisticated ML and DL models. The majority of current research ignores the dependability of probabilistic outputs and their consequences for decision support systems in favor of concentrating mostly on classification accuracy. Thus, creating prediction frameworks that are well-calibrated and aware of uncertainty continues to be a pertinent and important area of study in contemporary banking analytics.

Review of related work and problem statement. Machine learning-based methods considerably outperform conventional statistical models in capturing intricate and nonlinear consumer behavior patterns, according to recent research on bank customer churn prediction. Specifically, when applied to real banking datasets with high complexity and class imbalance, ensemble learning techniques have shown great discriminative power.

Random Forest and XGBoost consistently outperform single classifiers across various sampling strategies, with classification accuracies ranging from 87% to 96% and ROC-AUC values between 0.90 and 0.93, according to a thorough evaluation of several supervised learning models carried out on actual bank customer data [2]. Furthermore, robust and balanced performance under class-imbalanced situations, which are frequently seen in churn prediction problems, is indicated by reported F1-scores above 0.78. These findings demonstrate how well ensemble-based models can detect high-reliability, churn-prone clients in real-world banking settings.

Recent research has focused more on model transparency and probabilistic trustworthiness as essential prerequisites for practical implementation in financial decision-making systems, going beyond overall forecast accuracy. The optimized model obtained accuracy, precision, recall, F1-score, and AUC values all exceeding 0.95, with certain metrics reaching up to 0.97, according to an interpretable churn prediction framework based on XGBoost, imbalance handling, and SHAP-based explainability [6]. Additionally, important churn drivers like transactional intensity, customer tenure, and service usage frequency were found using feature contribution analysis utilizing SHAP. By combining transparent decision support with excellent predictive performance, this technique represents a practical development that helps banks understand not only which clients are likely to churn, but also why.

The excellent performance of classical ensemble models when improved through optimization approaches is further supported by recent empirical data. A Genetic Algorithm–optimized XGBoost (GA-XGBoost) model was used to investigate bank card customer churn on real transactional data. The results were extremely competitive, correctly identifying 452 out of 488 churn customers with a recall of 0.9262 and correctly classifying 2487 out of 2551 non-churn customers with a non-churn accuracy rate of 0.8760 [7]. Additionally, a ROC-AUC value of 0.9912, which shows almost perfect separation between churn and non-churn clients, verified the model's overall discriminative capacity. The use of genetic algorithms for XGBoost hyperparameter optimization with AUC as the fitness function and the incorporation of SHAP-based global and local explanations to lessen the black-box nature of ensemble models are the main innovations of this work. Nevertheless, the model does not specifically handle predictive uncertainty and depends on deterministic probability estimations, despite the high performance metrics.

While predictive performance measures have gradually increased, probability calibration, uncertainty awareness, and business-oriented evaluation criteria are still understudied in churn prediction research, according to recent analytical surveys and review studies. Even while churn mitigation solutions in banking often rely on risk thresholds and confidence levels rather than binary class labels alone, many previous research prioritize accuracy-based measurements, frequently ignoring the dependability of anticipated probability. Because of this, sophisticated machine learning models can provide overconfident forecasts, which would restrict their usefulness and possibly result in an ineffective use of retention resources.

The purpose of this work is to overcome the constraints noted in the literature regarding probability reliability and decision-oriented evaluation while examining the efficacy of ensemble-based machine learning models for bank customer churn prediction. The study specifically focuses on creating and evaluating churn prediction frameworks that combine robust discriminative performance with interpretable and well-calibrated probabilistic outputs, making them appropriate for risk-sensitive decision-making in banking settings. This work aims to close the gap between forecast accuracy and usefulness in real-world customer retention systems by fusing traditional ensemble models with uncertainty-aware learning techniques.

Research methods

Dataset description and preprocessing. A publicly accessible bank customer turnover dataset that includes the financial, behavioral, and demographic characteristics of retail banking clients is used for the experimental investigation. Customer age, credit score, account balance, tenure, quantity of items, activity status, and geography and gender data are among the factors included in the dataset. The target variable shows whether a customer has stayed active or left the bank (churned).

To avoid information leakage, unnecessary identifiers (customer ID, row number, and surname) are eliminated prior to model development. One-hot encoding is used to transform categorical characteristics, such as gender and geography, while standardizing numerical variables guarantees consistent feature scaling. To maintain the original churn distribution, the dataset is then divided into training and testing subsets using stratified sampling. In addition to ensuring data integrity, this preprocessing pipeline offers a solid basis for uncertainty-aware churn prediction modeling.

Machine learning models. An ensemble learning technique called Random Forest (RF) builds several decision trees using random feature selection and bootstrap sampling, then aggregates their predictions to enhance resilience and generalization performance. RF has been extensively used in banking analytics and customer churn prediction tasks, especially in class-imbalanced situations, because of its capacity to describe nonlinear connections and reduce variation through ensemble averaging [8].

By maximizing a regularized objective function, the gradient boosting framework XGBoost creates additive tree-based models that facilitate effective learning of intricate feature relationships. XGBoost is a cutting-edge method for churn prediction in financial datasets, where high predictive accuracy and stability are necessary, thanks to its scalability, integrated regularization, and potent discriminative capabilities [9].

By executing several stochastic forward passes during inference, neural networks in conjunction with Monte Carlo Dropout (MCDO) offer an uncertainty-aware learning paradigm. In risk-sensitive applications like banking decision-making systems, this method approximates Bayesian inference and allows for the estimate of predictive uncertainty in addition to point predictions [10].

Evaluation Metrics A number of performance and calibration metrics, such as the Area Under the Receiver Operating Characteristic Curve (AUC), F1-score, Brier score, Expected Calibration Error (ECE), and predictive uncertainty, were used to guarantee a thorough assessment of both discriminative performance and probabilistic reliability.

The model's overall discriminative capacity across all potential classification thresholds was assessed using the Area Under the Receiver Operating Characteristic Curve (AUC), which shows how effectively the model distinguishes between churn and non-churn customers independent of a set decision boundary [11]. Equation (1) defines the AUC metric as follows:

$$AUC = \int_0^1 TPR(t)\, dFPR(t) \qquad (1)$$

where the true positive rate (TPR) and false positive rate (FPR) are indicated, respectively.

The F1-score, a balanced performance metric that combines precision and recall, was used to alleviate the class imbalance frequently found in churn prediction datasets [12]. As stated in Eq. (2), the F1-score is calculated as the harmonic mean of precision and recall.

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \qquad (2)$$

The Brier score, which calculates the mean squared difference between expected probability and observed binary outcomes, was used to assess the accuracy of anticipated probabilities in addition to classification performance [13]. Better probabilistic accuracy and calibration are indicated by lower Brier score values. Equation (3) defines the Brier score:

$$Brier = \frac{1}{N}\sum_{i=1}^{N}(\widehat{p_\iota} - y_i)^2 \qquad (3)$$

where $\hat{p}_i$ represents the predicted probability and y_i denotes the true class label.

To further assess probability calibration quality, the Expected Calibration Error (ECE) was computed, measuring the discrepancy between predicted confidence levels and empirical accuracy across probability bins [14]. The ECE metric is defined as shown in Eq. (4):

$$ECE = \sum_{m=1}^{M}\frac{|B_m|}{N}|acc(B_m) - conf(B_m)| \qquad (4)$$

where B_m denotes the set of samples in the m-th probability bin, acc(B_m) is the empirical accuracy, and conf(B_m) is the average predicted confidence within that bin.

Monte Carlo Dropout was used to quantify predictive uncertainty for uncertainty-aware modeling by calculating the standard deviation of predicted probability derived from several stochastic forward passes during inference [10]. Eq. (5) illustrates how the prediction uncertainty is computed:

$$Uncertainty = \sqrt{\frac{1}{S}\sum_{s=1}^{S}(\hat{p}^{(s)} - \bar{p})^2} \qquad (5)$$

where S denotes the number of Monte Carlo forward passes and $\hat{p}^{(s)}$ represents the predicted probability from the s-th pass.

By allocating a gain of 500 AZN to each correctly identified churn client and a cost of 100 AZN to each false positive prediction, a profit-based metric was employed to quantify the economic impact of churn prediction. The projected profit was calculated using the formula defined in Eq. (6):

$$Profit = 500 \cdot TP - 100 \cdot FP \qquad (6)$$

Results and discussions

Predictive accuracy and calibration quality vary significantly amongst the used models, according to the performance comparison. As shown in Table, the Random

Forest (RF), XGBoost, and Neural Network (NN) combined with Monte Carlo Dropout (MCDO) each provided competitive but distinct outcomes. The Random Forest (Deep Ensemble) achieved a balanced performance, with an AUC of 0.842, an F1-score of 0.777, and an expected profit of 118,000 AZN. The model showed relatively low uncertainty (0.019) but a higher calibration error (ECE = 0.108), suggesting that while its classifications were stable, the confidence levels were somewhat overestimated.

Table – Comparative performance of models

Model	AUC	F1-Score	Brier	ECE	Uncertainty	Profit (AZN)
RF (Deep Ensemble)	0.842	0.777	0.130	0.108	0.019	118.000
XGBoost	0.833	0.794	0.127	0.088	0.031	111.900
NN + MCDO	0.863	0.631	0.103	0.026	0.059	111.500

With an estimated return of 111,900 AZN, the XGBoost model generated the highest F1-score (0.794) among all the models, with an AUC of 0.833. The dependability curve did, however, show a moderate calibration deviation (ECE = 0.088), suggesting minor discrepancies between expected probabilities and actual results.

The most reliable calibration performance and uncertainty representation were obtained by the Neural Network with Monte Carlo Dropout (NN + MCDO). Its probabilistic outputs were more in line with the real class frequencies, as seen by the AUC value of 0.863 and the lowest calibration error (ECE = 0.026). The MCDO model demonstrated a more reliable confidence estimation and realistic uncertainty distribution (0.059), even though it had a slightly lower F1-score (0.631) than XGBoost. This characteristic is crucial for financial prediction systems because it improves interpretability and encourages more risk-aware decision-making.

Overall, the comparison shows that the NN + MCDO technique showed better calibration strength, less overconfidence, and more consistent uncertainty quantification while Random Forest and XGBoost produced great discriminative power. Because of these features, it is especially appropriate for financial settings that demand accurate probabilistic forecasts.

The dependability curve for the Neural Network with Monte Carlo Dropout (NN + MCDO) is shown in Figure Perfect calibration is represented by the diagonal dashed line, where the estimated probability and actual event frequencies should coincide. The model's confidence levels are in good agreement with observed results, as evidenced by the NN + MCDO curve's strong adherence to this reference line across the majority of probability bins.

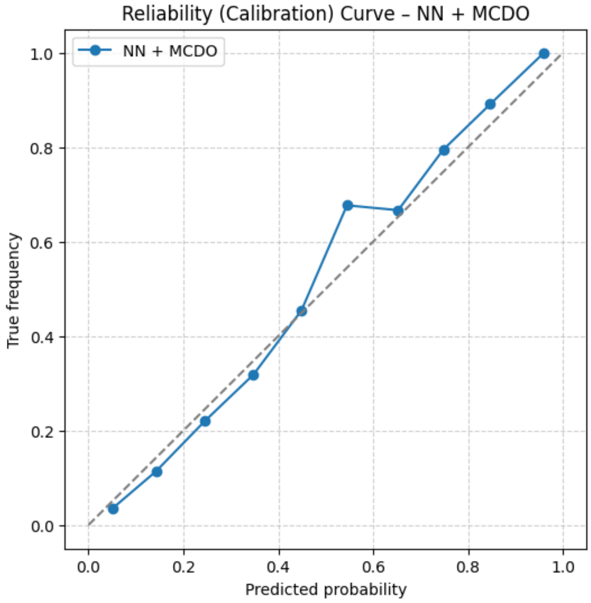

Figure – Reliability (Calibration) curve of the NN + MCDO model

This behavior demonstrates that the MCDO-based neural network minimizes overconfidence and enhances interpretability while producing extremely dependable probabilistic outputs. The low calibration error (ECE = 0.026) shown in Table 6 is reflected in the minimal difference between expected and true frequencies.

Conclusion

The usefulness of uncertainty-aware machine learning models for forecasting customer attrition in the banking industry was investigated in this study. As evidenced by improvements in calibration error (ECE decreased to 0.026) and probabilistic alignment (AUC reaching 0.863), the results demonstrated that combining probabilistic reasoning and calibration measures improves both the interpretability and reliability of prediction outcomes.

The ensemble-based methods guaranteed high predictive strength and financial profitability, achieving F1-scores up to 0.794 and expected profits of 118,000 AZN, while the neural network with Monte Carlo Dropout produced the most stable and well-calibrated probabilistic outputs (ECE = 0.026, uncertainty = 0.059). These results highlight the importance of a model's dependability in addition to its accuracy, particularly when forecasts have an impact on actual financial decisions.

Future research could enable adaptive, risk-sensitive, and customer-centric decision systems in digital banking by expanding uncertainty quantification techniques to hybrid and federated learning frameworks.

Conflict of Interests

The author declares there is no conflict of interests related to the publication of this article.

REFERENCES

1. **Tran H.D., Le N., Nguyen V.H.** Customer churn prediction in the banking sector using machine learning-based classification models. Interdisciplinary Journal of Information, Knowledge & Management, 18, 87–105, 2023. https://doi.org/10.28945/5086
2. **Ashraf R.** Bank customer churn prediction using machine learning framework. Journal of Applied Finance & Banking, 14(4), 1–5, 2024.
3. **Badalova A.N., Guliyeva S.H.** Application of Machine Learning Methods for Classification of Agricultural Crops. Herald of the Azerbaijan Engineering Academy, 14(2), 106–116, 2022. https://doi.org/10.52171/2076-0515_2022_14_02_106_116
4. **Imani M., Joudaki M., Beikmohammadi A., Arabnia H.R.** Customer churn prediction: A

systematic review of recent advances, trends, and challenges in machine learning and deep learning. Machine Learning and Knowledge Extraction, 7(3), 105, 2025. https://doi.org/10.3390/make7030105

5. **Xu X., Kou G., Ergu D.** Profit-based uncertainty estimation with application to credit scoring. European Journal of Operational Research, 325(2), 303–316, 2025. https://doi.org/10.1016/j.ejor.2025.03.007
6. **Li Y., Yan K.** Prediction of bank credit customers churn based on machine learning and interpretability analysis. Data Science in Finance and Economics, 5(1), 19–34, 2025. https://doi.org/10.3934/DSFE.2025002
7. **Peng K., Peng Y., Li W.** Research on customer churn prediction and model interpretability analysis. PLoS ONE, 18(12), e0289724, 2023. https://doi.org/10.1371/journal.pone.0289724
8. **Breiman, L.** Random forests. Machine Learning, 45(1), 5–32, 2001. https://doi.org/10.1023/A:1010933404324
9. **Chen, T., & Guestrin, C.** XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794, 2016. https://doi.org/10.1145/2939672.2939785
10. **Gal, Y., & Ghahramani, Z.** Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning (ICML), 1050–1059, 2016
11. **Fawcett, T.** *An introduction to ROC analysis*. Pattern Recognition Letters, 27(8), 861–874, 2006. **DOI:** https://doi.org/10.1016/j.patrec.2005.10.010
12. **Powers, D.M.W.** Evaluation: From precision, recall and F-measure to ROC. Journal of Machine Learning Technologies, 2(1), 37–63, 2011
13. **Brier, G.W.** Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1), 1–3, 1950
14. **Guo, C., Pleiss, G., Sun, Y., & Weinberger, K.Q.** On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning (ICML), 1321–1330, 2017