QUBAHAN JOURNAL OF COACHING AND SPORTS SCIENCES
VOL. 4, NO. 3, Sep 2026
https://doi.org/10.48161/qcsj.v4n2a88

Cross-Domain Uncertainty Evaluation in Predictive Analytics: Credit Default and Customer Churn

Heitor Gabriel da Silva Valente[1*], Etelka Éva Katits[2]

[1] Department of Sports Management and Economics, Chengdu Sport University, Chengdu, China;
[2] Department of Finance and Management, University of Pannonia, Zalaegerszeg, Hungary;

* **Corresponding author:** gabriheitor@gmail.com.

ABSTRACT: This study develops and evaluates a cross-domain, probability-based framework for evaluating uncertainty in predictive analytics. Although predictive models widely support decision-making, many models are evaluated mainly using accuracy-based indicators that do not show whether predicted probabilities are reliable, calibrated, or suitable for decision support. The proposed framework integrates probabilistic classification, Bayesian band-level estimation, calibration diagnostics, bootstrap validation, scoring rules, and probability-based decision bands. Two public benchmark datasets were used: a credit-card default dataset with 30,000 observations as the main financial-risk case and a customer-churn dataset with 7,043 observations as a secondary marketing-analytics validation case. Logistic regression, random forest, and gradient boosting models were compared using conventional metrics and uncertainty-aware indicators, including the area under the receiver operating characteristic curve (ROC-AUC), accuracy, precision, recall, F1-score, Brier score, log loss, expected calibration error, maximum calibration error, bootstrap intervals, calibration curves, and Bayesian credible intervals for observed event rates within probability bands. The results show that gradient boosting provided the strongest probability reliability in both empirical cases, with a Brier score of 0.1380 and expected calibration error of 0.0179 in credit default prediction and a Brier score of 0.1407 and expected calibration error of 0.0346 in customer churn prediction. Decision-band analysis showed that borderline cases had observed event rates near 54.33% for credit default and 52.43% for customer churn, confirming that uncertainty bands can separate confident predictions from review-needed cases. The study advances trustworthy predictive analytics by showing that model evaluation should combine predictive accuracy with probability reliability, calibration quality, and uncertainty communication.

Keywords: predictive analytics; uncertainty evaluation; probability framework; calibration; Bayesian estimation.

I. INTRODUCTION

Predictive analytics has become a central instrument for decision-making in finance, healthcare, marketing, education, industrial monitoring, and public-service planning. In these settings, a predicted class or expected outcome is often treated as sufficient evidence for action. A bank may approve or decline a credit request, a marketing team may decide whether to retain a customer, and an industrial analyst may decide whether a system requires intervention. Yet a prediction's practical value depends not only on whether the predicted label is correct. It also depends on the reliability of the assigned probability, how uncertain the model is in similar cases, and whether it distinguishes between confident and borderline cases. A model that is accurate on average may still be unsuitable for decision support if its probability estimates are poorly calibrated or if it presents uncertain predictions as confident outputs.

The problem becomes more important when predictive systems are used in environments where decisions carry costs. In credit-risk analytics, a false low-risk classification may expose an institution to

financial loss, while an overly conservative high-risk classification may reduce fair access to credit. In customer-retention analytics, a poorly calibrated churn probability may lead to inefficient incentives, unnecessary retention expenditure, or missed intervention opportunities. In both domains, therefore, interpret a predictive model's output probabilistically. Decision-makers need to know whether a predicted probability of 0.20 actually corresponds to a low observed event rate, whether a predicted probability near 0.50 should be treated as borderline, and whether the model behaves similarly across different empirical contexts.

This study addresses that need by proposing and testing a probability-based framework for uncertainty evaluation in predictive analytics. The framework does not replace conventional predictive evaluation; rather, it extends it. Accuracy, precision, recall, F1-score, and AUC remain useful for measuring classification performance, but they are not enough for judging probability reliability. The proposed framework adds probability scoring, calibration diagnostics, bootstrap intervals, Bayesian band-level credible intervals, and decision-band analysis. These components allow evaluation of a predictive model in terms of both performance and uncertainty.

The empirical design uses two public benchmark datasets. The credit-card default dataset is used as the main case because credit default is naturally a probabilistic risk problem. The Telco customer churn dataset is used as a secondary validation case because churn prediction is a common applied predictive-analytics task in marketing and customer-retention research. The datasets are not merged because they describe different populations, variables, and decision contexts. Instead, we use them as cross-domain evidence to assess whether the proposed uncertainty-evaluation framework remains meaningful across two separate predictive tasks.

The study contributes to predictive analytics in three ways. First, it offers a structured framework that connects probability modeling, calibration, uncertainty intervals, and decision bands. Second, it demonstrates the framework empirically using two datasets from different application domains. Third, it shows that the most useful predictive model is not necessarily the one with the highest discrimination score, but the one that provides reliable probabilities for decision support. The central argument is that predictive analytics should not be evaluated only by whether it predicts correctly, but also by whether it communicates uncertainty responsibly.

II. RESEARCH AIM, RESEARCH QUESTIONS, AND HYPOTHESES

This study aims to develop and empirically evaluate a probability-based framework for uncertainty evaluation in predictive analytics. The framework aims to show how probability estimates, calibration quality, scoring rules, bootstrap intervals, Bayesian credible intervals, and decision bands can be integrated to evaluate the reliability of predictive outputs in addition to conventional accuracy-based performance.

The study is guided by the following research questions. First, how can probability-based methods be integrated into a structured framework for uncertainty evaluation in predictive analytics? Second, how do conventional predictive-performance metrics compare with uncertainty-aware metrics when evaluating credit default and customer churn models? Third, which model provides the most reliable probability estimates across the two empirical datasets? Fourth, how can uncertainty bands help decision-makers separate confident predictions from borderline or high-uncertainty cases?

The study tests three hypotheses.

- H1: Models evaluated with probability-based uncertainty metrics provide more decision-relevant evidence than models evaluated only by conventional classification metrics.
- H2: Calibration and probability scoring reveal reliability differences that are not fully captured by AUC, accuracy, precision, recall, or F1-score.
- H3: A cross-domain uncertainty-evaluation framework can identify consistent decision bands across finance and marketing predictive-analytics tasks.

The figure summarizes the study contribution across methodological, empirical, practical, and evaluation dimensions.

III. STUDY CONTRIBUTION

The study makes four main contributions. First, it proposes a probability-based framework that integrates conventional predictive evaluation with uncertainty-aware diagnostics. Second, it demonstrates the framework using two public datasets from different applied domains. Third, it shows that decision bands can convert probability outputs into practical risk categories. Fourth, it adds Bayesian credible intervals to band-level event rates, allowing uncertainty to be reported even after the model has produced probability classifications.

IV. LITERATURE REVIEW

Predictive modeling research emphasizes that applied prediction should be treated as a complete workflow rather than as a single algorithmic exercise. Kuhn and Johnson [12] described applied predictive modeling as a process that requires data preparation, resampling, feature treatment, model comparison, and validation. Bayesian statistical work similarly established that uncertainty should be represented through probability distributions rather than hidden behind point estimates [7, 11]. For the present study, these foundations are important because the framework is built around predicted probabilities, posterior uncertainty, and decision-oriented validation.

Transparent reporting and validation have become central requirements in prediction-model research. Collins et al. [5] introduced reporting standards for prediction-model studies, while Moons et al. [18] expanded the explanation of model-development and validation components. These studies show that predictive models should not be judged only by headline accuracy, because useful prediction also depends on data description, validation design, calibration, and intended use. In the present manuscript, this logic supports a clear distinction between conventional classification performance and uncertainty-aware probability evaluation.

Probability calibration has developed into a major concern in predictive analytics. Naeini et al. [19] proposed Bayesian binning into quantiles as a method for improving probability calibration in binary classifiers. Gal and Ghahramani [6] connected model uncertainty with approximate Bayesian inference in neural prediction, while Ribeiro et al. [22] emphasized that users need interpretable evidence for model outputs. Together, these studies indicate that a prediction is useful only if its probability estimate can be interpreted and communicated clearly.

Bayesian and approximate Bayesian methods provide a mathematical basis for uncertainty-aware prediction. Blei et al. [3] reviewed variational inference as a scalable method for approximating difficult posterior distributions, and Carpenter et al. [4] demonstrated how probabilistic programming can support applied Bayesian modeling. These contributions matter to the present framework because they treat uncertainty estimation as a formal part of modeling rather than a secondary descriptive step.

Calibration research has also shown that high discrimination does not guarantee reliable probability estimates. Guo et al. [8] demonstrated that modern classifiers can produce poorly calibrated probabilities despite strong classification performance. Kendall and Gal [10] separated aleatoric and epistemic uncertainty, showing that uncertainty has different sources and should not be reduced to a single error rate. Lakshminarayanan et al. [15] further demonstrated that ensemble approaches can improve uncertainty estimation. These findings motivate the present study's use of calibration curves, Brier score, log loss, and uncertainty bands.

Model evaluation in probabilistic and Bayesian settings has increasingly moved toward out-of-sample predictive assessment. Vehtari et al. [28] showed that Bayesian model comparison can be based on predictive accuracy estimates that account for uncertainty in fitted models. Leibig et al. [16] showed the practical value of uncertainty information in sensitive detection tasks. Kuleshov et al. [13] added that uncertainty intervals themselves must be calibrated, since nominal intervals may not match observed coverage. This literature supports the paper's emphasis on whether predicted probabilities and intervals correspond to empirical outcomes.

Several studies have tried to make calibration directly measurable. Kumar et al. [14] addressed calibration as an objective that can be optimized, and Yao et al. [30] developed stacking for Bayesian predictive distributions. Ovadia et al. [21] showed that uncertainty estimates may deteriorate under distribution shift, which is especially relevant to real-world decision systems. Riley et al. [23] emphasized that prediction

models require adequate sample size and outcome events to avoid unstable estimates. These studies together justify using two large datasets and reporting both discrimination and uncertainty indicators.

Further work confirms that uncertainty evaluation is sensitive to model design and assessment choices. Thulasidasan et al. [25] linked training strategy to improved calibration, while Vaicenavicius et al. [26] developed a formal framework for evaluating calibration in classification. Van Calster et al. [27] argued that calibration is a major weakness in predictive analytics, especially when models are used for decision support. Nixon et al. [20] warned that calibration metrics can differ and should be interpreted carefully. These findings support the present study's use of multiple uncertainty-aware metrics rather than relying on a single calibration statistic.

Applied prediction studies also show that poor uncertainty evaluation can weaken decision support in high-stakes contexts. Mehrtash et al. [17] studied confidence calibration and predictive uncertainty in medical image segmentation, while Ashukha et al. [2] highlighted pitfalls in uncertainty estimation and ensemble evaluation. Wynants et al. [29] showed that prediction models can become unreliable when validation and reporting are weak. Angelopoulos and Bates [1] introduced conformal prediction as a way to communicate distribution-free uncertainty. These studies show why communicating uncertainty is practically necessary in applied predictive analytics.

Finally, Hullermeier and Waegeman [9] provided a clear conceptual distinction between aleatoric and epistemic uncertainty, which helps explain why decision-makers need to know whether uncertainty comes from noisy data or limited model knowledge. Senge et al. [24] further emphasized that reliable classification should quantify uncertainty sources rather than only return class labels. The present study builds on this literature by proposing a cross-domain probability-based framework and by testing it with credit default and customer churn prediction models.

Table 1. Literature review comparison of probability-based uncertainty evaluation studies.

No.	Study	Focus	Main contribution	Use in this paper
1	Angelopoulos and Bates (2021)	Conformal prediction	Distribution-free uncertainty quantification can provide valid predictive sets.	Supports interval-oriented uncertainty communication.
2	Ashukha et al. (2020)	Pitfalls of uncertainty estimation	Some uncertainty methods can appear strong in-domain but fail under evaluation changes.	Supports careful empirical validation.
3	Blei et al. (2017)	Variational inference	Approximate Bayesian inference is useful when computing the exact posterior is difficult.	Supports scalable uncertainty estimation.
4	Carpenter et al. (2017)	Probabilistic programming	Probabilistic modelling environments support Bayesian inference for applied research.	Supports reproducible Bayesian implementation.
5	Collins et al. (2015)	Prediction-model reporting	Transparent reporting is necessary for judging the usefulness of prediction models.	Guides manuscript reporting and validation clarity.
6	Gal and Ghahramani (2016)	Model uncertainty in deep models	Dropout can be interpreted as approximate Bayesian inference for uncertainty estimation.	Positions uncertainty as a central modeling issue.
7	Gelman et al. (2013)	Bayesian data analysis	Bayesian inference represents uncertainty through probability distributions and posterior estimation.	Justifies the Bayesian component of the framework.
8	Guo et al. (2017)	Calibration of neural classifiers	High predictive accuracy does not guarantee well-calibrated probabilities.	Motivates calibration curves and ECE.

No.	Study	Focus	Main contribution	Use in this paper
9	Hullermeier and Waegeman (2021)	Aleatoric and epistemic uncertainty	A clear distinction between data noise and model uncertainty is needed.	Supports conceptual framework.
10	Kendall and Gal (2017)	Aleatoric and epistemic uncertainty	Separate different uncertainty sources for reliable predictive systems.	Supports uncertainty taxonomy.
11	Kruschke (2014)	Bayesian estimation and interpretation	Credible intervals provide interpretable uncertainty statements for model parameters and probabilities.	Supports posterior probability interpretation.
12	Kuhn and Johnson (2013)	Predictive modeling workflow	Systematic model training and validation are essential for applied prediction.	Supports the empirical modelling pipeline.
13	Kuleshov et al. (2018)	Calibrated regression uncertainty	Uncertainty intervals require calibration, not only interval construction.	Supports interval-calibration logic.
14	Kumar et al. (2018)	Trainable calibration measures	Incorporates calibration error into the learning objective.	Supports calibration as an optimization goal.
15	Lakshminarayanan et al. (2017)	Ensemble uncertainty	Deep ensembles can provide strong predictive uncertainty estimates.	Supports ensemble-based uncertainty thinking.
16	Leibig et al. (2017)	Uncertainty in disease detection	Uncertainty estimates can improve reliability in sensitive predictive applications.	Supports decision-support relevance.
17	Mehrtash et al. (2020)	Uncertainty in medical segmentation	Model confidence may be inaccurate without calibration and uncertainty diagnostics.	Supports uncertainty-aware validation in high-stakes domains.
18	Moons et al. (2015)	Prediction-model explanation and elaboration	Validation, missing-data handling, and model-use context should be clearly reported.	Supports reproducible methodology.
19	Naeini et al. (2015)	Bayesian probability calibration	Bayesian binning improves calibration of probabilistic classifiers.	Supports post-model probability evaluation.
20	Nixon et al. (2019)	Calibration measurement in deep learning	Calibration metrics vary and should be interpreted carefully.	Supports multi-metric uncertainty assessment.
21	Ovadia et al. (2019)	Uncertainty under dataset shift	Uncertainty and calibration can degrade when test conditions differ from training data.	Supports cross-domain validation.
22	Ribeiro et al. (2016)	Model interpretability	Local explanations help users understand individual model outputs.	Connects uncertainty with decision transparency.
23	Riley et al. (2019)	Sample-size adequacy for prediction models	Prediction studies require enough events and participants to reduce overfitting.	Supports dataset-size justification.
24	Senge et al. (2014)	Reliable classification uncertainty	Classifiers can distinguish aleatoric and epistemic uncertainty rather than only return labels.	Supports uncertainty-aware classification.
25	Thulasidasan et al. (2019)	Calibration through mixup training	Training strategy can affect calibration and uncertainty quality.	Supports calibration sensitivity.

No.	Study	Focus	Main contribution	Use in this paper
26	Vaicenavicius et al. (2019)	Calibration evaluation theory	Calibration requires formal probability-based evaluation and careful visualization.	Supports reliability diagrams and ECE.
27	Van Calster et al. (2019)	Calibration in predictive analytics	Poor calibration is a major weakness in applied predictive models.	Directly supports the paper's rationale.
28	Vehtari et al. (2017)	Bayesian model evaluation	Predictive accuracy can be evaluated through out-of-sample Bayesian criteria.	Supports validation beyond a single train-test split.
29	Wynants et al. (2020)	Critical appraisal of prediction models	Poor reporting and validation can make prediction models unreliable.	Supports methodological discipline.
30	Yao et al. (2018)	Bayesian predictive stacking	Combining predictive distributions can improve Bayesian prediction in open settings.	Supports probability-distribution comparison.

V. RESEARCH GAP

The reviewed literature provides strong theoretical and methodological support for uncertainty evaluation, but three gaps remain. First, many studies focus on calibration, Bayesian inference, or uncertainty intervals separately, while few integrate these components into a single applied decision-support framework. Second, many empirical studies focus on a single domain, making it difficult to assess whether a probability-based uncertainty framework transfers across predictive contexts. Third, conventional predictive-analytics studies often report AUC and accuracy but do not explain how decision-makers should treat borderline cases. This study addresses these gaps by developing a unified framework and applying it to two different binary predictive tasks: credit default and customer churn.

VI. PROPOSED PROBABILITY-BASED FRAMEWORK

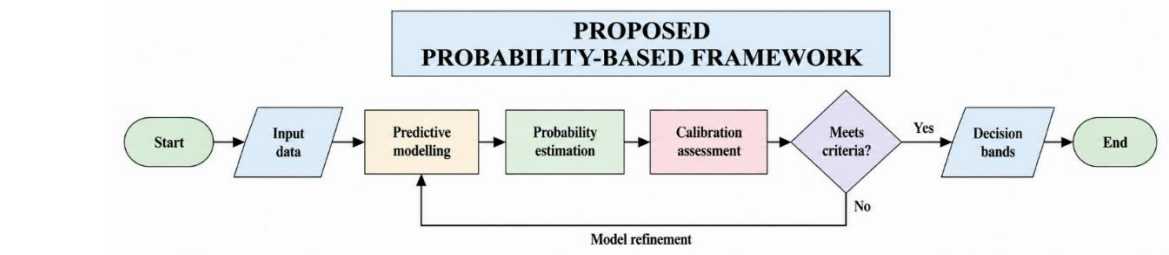


FIGURE 1. Proposed Method.

The framework connects predictive modelling, probability estimation, calibration assessment, Bayesian uncertainty intervals, and decision-support interpretation into one reproducible evaluation structure.

VII. METHODOLOGY

1. RESEARCH DESIGN

The study uses a quantitative empirical research design based on supervised binary classification. The goal is not only to identify the most accurate model but also to evaluate whether predicted probabilities are reliable enough for decision support. The empirical analysis follows a cross-domain validation design. The first case examines credit-card default prediction, and the second case examines customer churn prediction.

Both tasks are suitable because the outcome is binary, the decision context is practical, and the model output can be interpreted as an event probability.

2. DATA SOURCES

The main dataset is the UCI Credit Card Default dataset downloaded through Kaggle. It contains 30,000 client-level observations and 25 variables, including an identifier, demographic attributes, repayment status indicators, bill amounts, payment amounts, and the binary outcome default.payment.next.month. The secondary dataset is the Telco Customer Churn dataset, also downloaded from Kaggle. It contains 7,043 customer observations and 21 variables, including customer service characteristics, contract information, billing information, and the binary outcome Churn. We did not merge the two datasets because they refer to different domains, populations, variables, and outcome meanings. Instead, we used them as separate empirical cases within a common uncertainty-evaluation framework.

3. DATA PREPARATION

For the credit dataset, the identifier variable was removed and the target variable default.payment.next.month was coded as the positive outcome. We treated repayment-status and demographic variables as categorical because their numeric codes represent categories or ordinal status indicators rather than continuous quantities. For the Telco dataset, customerID was removed and Churn was recoded as a binary target. The TotalCharges field contained 11 blank values, which were converted to missing numeric values and imputed during preprocessing. Categorical predictors were encoded within the modelling pipeline, and numeric variables were imputed and scaled where required. We split each dataset into 70% training and 30% testing sets using stratified sampling to preserve the positive-outcome distribution.

4. MODELS

Three predictive models were estimated for both datasets. Logistic regression was used as an interpretable baseline. Random forest was used to capture nonlinear relationships through ensemble learning. Gradient boosting was used as a strong boosting-based classifier. We evaluated the models on the same test sets using conventional and uncertainty-aware indicators. We selected the uncertainty model with the strongest combined probability reliability, emphasizing the Brier score, expected calibration error, and log loss rather than accuracy alone.

5. PROBABILITY-BASED UNCERTAINTY EVALUATION

The proposed framework begins with probabilistic binary prediction. For each case i, a model estimates the event probability as:

$$\hat{p}_i = P(Y_i = 1 \mid X_i = x_i)$$

For logistic regression, the estimated event probability is obtained from the logistic transformation:

$$\hat{p}_i = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})]}$$

For ensemble prediction, the probability estimate can be represented as the average predicted probability across T learners:

$$\hat{p}_i = \frac{1}{T}\sum_{t=1}^{T} \hat{p}_{it}$$

The Brier score measures the squared distance between predicted probabilities and observed binary outcomes:

$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{n} (\hat{p}_i - y_i)^2$$

The log-loss score penalizes confident but incorrect probability estimates:

$$\mathrm{LL} = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i)]$$

Expected calibration error compares predicted probability and observed event frequency across B probability bins:

$$\mathrm{ECE} = \sum_{b=1}^{B}\frac{n_b}{n}|\mathrm{acc}(b) - \mathrm{conf}(b)|$$

Bootstrap uncertainty intervals are estimated by repeated resampling of the test set:

$$\mathrm{CI}_{95\%}(\theta) = \left[Q_{0.025}(\hat{\theta}^*), Q_{0.975}(\hat{\theta}^*)\right]$$

For Bayesian decision-band evaluation, the observed event rate in each probability band is represented using a beta-binomial update:

$$\theta_b \mid y_b, n_b \sim \mathrm{Beta}(\alpha + y_b, \beta + n_b - y_b)$$

where y_b is the number of observed events in band b, n_b is the number of cases in the band, and we use a uniform Beta(1,1) prior. The 95% credible interval for each band is taken from the posterior beta distribution. This step allows the empirical event rate in each risk band to be interpreted with uncertainty rather than as a fixed percentage.

6. VALIDATION PROCEDURE

The analysis reports conventional metrics and uncertainty-aware metrics. Conventional metrics include accuracy, precision, recall, F1-score, and AUC. Uncertainty-aware metrics include Brier score, log loss, expected calibration error, maximum calibration error, calibration curves, bootstrap confidence intervals, and Bayesian credible intervals for decision-band event rates. This combination allows the study to evaluate both discrimination and probability reliability. Lower Brier score, lower log loss, and lower calibration error indicate stronger probability quality. AUC remains useful, but it is interpreted as a discrimination metric rather than as direct evidence of probability reliability.

7. VALIDITY, RELIABILITY, AND REPRODUCIBILITY

Stratified data splitting, consistent preprocessing pipelines, and evaluation on held-out test sets supported internal validity. We supported reliability by applying the same framework to two different domains. We strengthened reproducibility by using public datasets, transparent preprocessing rules, fixed random seeds, and standard metrics. The study does not claim causal effects. The results are interpreted as predictive and probabilistic evidence only.

Table 2. Dataset summary and empirical design.

Dataset	Rows	Predictors after ID/target removal	Positive outcome rate (%)	Test-set rows	Test positive rate (%)	Numeric predictors	Categorical predictors	Best uncertainty model
Credit default	30000	23	22.12	9000	22.12	14	9	Gradient boosting
Customer churn	7043	19	26.54	2113	26.55	4	15	Gradient boosting

Table 2 confirms that the credit default dataset is the stronger main case because of its larger sample size, while the customer churn dataset provides useful secondary validation from a different applied domain.

Table 3. Predictive performance and uncertainty-aware model comparison.

Dataset	Model	Accuracy	Precision	Recall	F1-score	AUC	Brier score	Log loss	ECE	MCE	Training time (s)
Credit default	Logistic regression	0.7749	0.4923	0.564	0.5257	0.7636	0.1851	0.5653	0.2133	0.2676	0.65

Dataset	Model	Accuracy	Precision	Recall	F1-score	AUC	Brier score	Log loss	ECE	MCE	Training time (s)
Credit default	Random forest	0.7762	0.4952	0.6017	0.5433	0.7747	0.1806	0.5481	0.2065	0.2712	1.7
Credit default	Gradient boosting	0.8183	0.6749	0.3451	0.4566	0.764	0.138	0.4405	0.0179	0.0592	3.37
Customer churn	Logistic regression	0.744	0.5114	0.7968	0.623	0.8444	0.1673	0.4939	0.1504	0.3115	0.08
Customer churn	Random forest	0.7501	0.5191	0.8004	0.6297	0.8431	0.1619	0.4844	0.1438	0.2517	0.67
Customer churn	Gradient boosting	0.7932	0.689	0.4029	0.5084	0.8367	0.1407	0.4341	0.0346	0.1574	0.36

Table 3 shows that accuracy and AUC alone do not provide a complete evaluation. Gradient boosting had the best probability reliability in both datasets because it produced the lowest Brier score and much lower calibration error than the other models.

Table 4. Bootstrap uncertainty intervals for best probability models.

Dataset	Best model	Metric	2.5%	97.5%
Credit default	Gradient boosting	AUC	0.7515	0.7745
Credit default	Gradient boosting	Brier score	0.1339	0.1422
Credit default	Gradient boosting	Log loss	0.4307	0.4513
Customer churn	Gradient boosting	AUC	0.8192	0.8551
Customer churn	Gradient boosting	Brier score	0.1333	0.148
Customer churn	Gradient boosting	Log loss	0.4156	0.4519

Table 4 shows that the uncertainty intervals are relatively narrow in the larger credit dataset and wider in the smaller churn dataset, which is consistent with sample-size expectations.

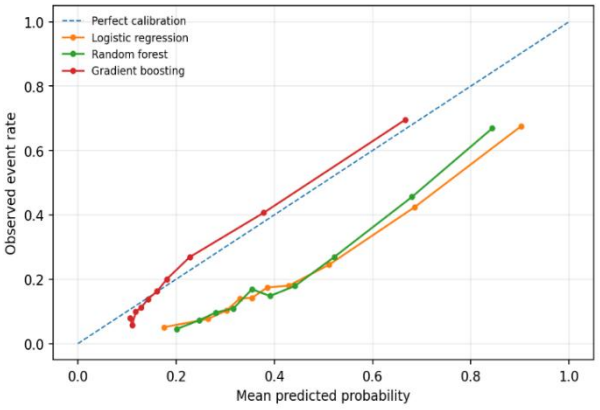


FIGURE 2. Calibration curve for the credit default models.

The calibration curve indicates that gradient boosting is closer to the ideal diagonal line than the baseline models, supporting its selection as the main uncertainty model for credit default.

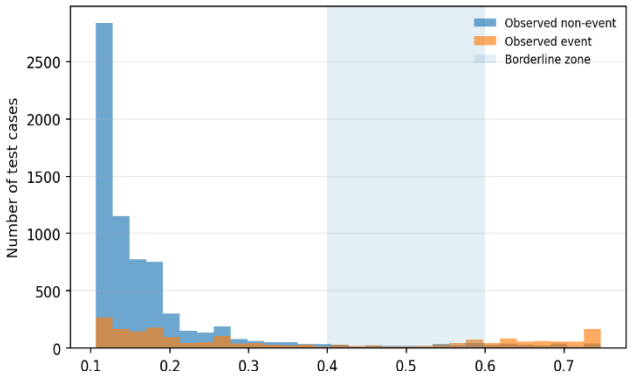


FIGURE 3. Predicted probability distribution and borderline zone for credit default.

The distribution shows that many credit cases fall in low-risk probability ranges, while a smaller but important group appears in the borderline zone requiring careful review.

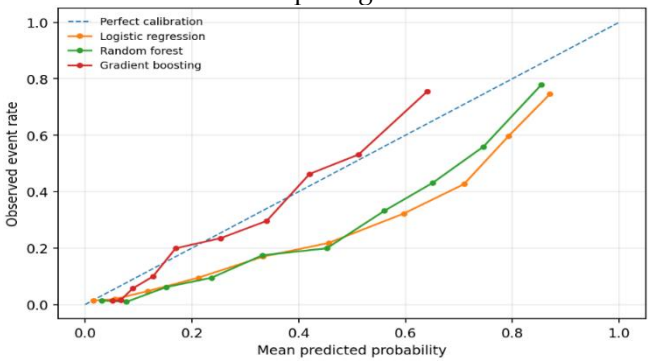


FIGURE 4. Calibration curve for the customer churn models.

The churn calibration curve confirms that uncertainty-aware evaluation is useful beyond finance, because probability reliability also varies across marketing prediction models.

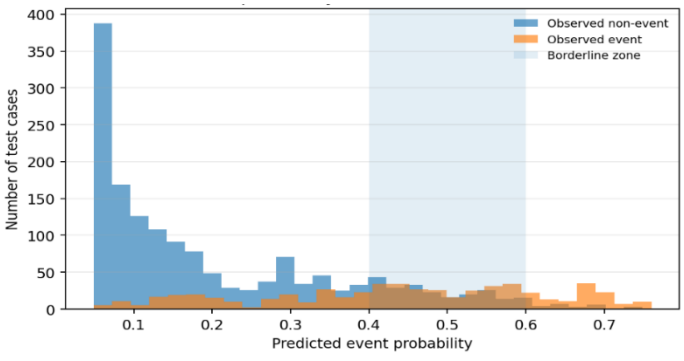


FIGURE 5. Predicted probability distribution and borderline zone for customer churn.

The distribution shows a visible borderline region, where automated classification alone is less reliable for retention decisions.

Table 5. Probability decision bands and observed event rates.

Dataset	Best model	Band	Probability range	Cases	Share (%)	Observed event rate (%)	Mean predicted probability (%)
Credit default	Gradient boosting	Low risk	0.00-0.20	6450	71.67	12.47	13.64
Credit default	Gradient boosting	Moderate risk	0.20-0.40	1361	15.12	30.64	26.92
Credit default	Gradient boosting	Borderline/uncertain	0.40-0.60	462	5.13	54.33	51.81
Credit default	Gradient boosting	High risk	0.60-0.80	727	8.08	71.39	68.18
Credit default	Gradient boosting	Very high risk	0.80-1.00	0	0		
Customer churn	Gradient boosting	Low risk	0.00-0.20	1070	50.64	7.76	10.17
Customer churn	Gradient boosting	Moderate risk	0.20-0.40	451	21.34	27.94	30.39
Customer churn	Gradient boosting	Borderline/uncertain	0.40-0.60	452	21.39	52.43	49.1
Customer churn	Gradient boosting	High risk	0.60-0.80	140	6.63	82.14	66.96
Customer churn	Gradient boosting	Very high risk	0.80-1.00	0	0		

The table shows that observed event rates increase consistently from low-risk bands to high-risk bands, indicating that probability bands can communicate uncertainty in a practical decision-support format.

Table 6. Bayesian credible intervals for event rates within probability bands.

Dataset	Band	Cases	Observed events	Observed event rate (%)	Bayesian posterior mean (%)	95% credible interval lower (%)	95% credible interval upper (%)
Credit default	Low risk	6450	804	12.47	12.48	11.68	13.29
Credit default	Moderate risk	1361	417	30.64	30.67	28.25	33.14
Credit default	Borderline/uncertain	462	251	54.33	54.31	49.77	58.82
Credit default	High risk	727	519	71.39	71.33	68	74.55
Customer churn	Low risk	1070	83	7.76	7.84	6.3	9.52
Customer churn	Moderate risk	451	126	27.94	28.04	24	32.26
Customer churn	Borderline/uncertain	452	237	52.43	52.42	47.83	57
Customer churn	High risk	140	115	82.14	81.69	74.95	87.59

Table 6 adds Bayesian uncertainty to the observed event rates. Wider intervals in smaller bands show why borderline and high-risk decisions should be interpreted with caution rather than as fixed percentages.

VIII. RESULTS AND FINDINGS

1. CREDIT DEFAULT FINDINGS

The credit default case shows that probability-aware evaluation changes how we interpret model performance. Logistic regression achieved an AUC of 0.7636 but had a Brier score of 0.1851 and expected calibration error of 0.2133. Random forest improved AUC slightly to 0.7747 but still had a high expected calibration error of 0.2065. Gradient boosting had similar discrimination (AUC = 0.7640) but produced a much better Brier score (0.1380) and expected calibration error (0.0179). This means we did not select gradient boosting because it had the highest AUC, but because it produced more reliable probabilities.

The decision-band analysis also supports the usefulness of uncertainty evaluation. In the credit dataset, 71.67% of test cases fell into the low-risk band, and the observed default rate in this group was 12.47%. Moderate-risk cases had an observed default rate of 30.64%, while borderline cases had 54.33%. High-risk cases had an observed default rate of 71.39%. These results show that the probability bands are meaningful because observed default rates increase as predicted probability bands increase.

2. CUSTOMER CHURN FINDINGS

The customer churn case provides a second validation of the framework. Logistic regression had the highest AUC (0.8444) but a Brier score of 0.1673 and an expected calibration error of 0.1504. Random forest had similar discrimination (AUC = 0.8431), but its expected calibration error remained high at 0.1438. Gradient boosting had a slightly lower AUC of 0.8367 but produced a much better Brier score of 0.1407 and expected calibration error of 0.0346. This confirms the study's main argument: the best model for uncertainty-aware decision support is not necessarily the one with the highest AUC.

Decision-band analysis showed that 50.64% of churn test cases were classified as low risk, with an observed churn rate of 7.76%. Moderate-risk cases had an observed churn rate of 27.94%, borderline cases 52.43%, and high-risk cases 82.14%. These results show that probability bands help identify groups that require different decision strategies. Low-risk cases may not require intervention, high-risk cases may justify strong retention action, and borderline cases may require additional information before intervention.

3. CROSS-DOMAIN INTERPRETATION

Across both datasets, gradient boosting produced the best uncertainty profile because it combined acceptable discrimination with substantially stronger calibration and a lower Brier score. The credit and churn cases also showed similar band-level patterns. In both datasets, borderline cases had observed event rates close to 50%, which confirms that these cases are genuinely uncertain rather than misclassified. The results therefore support the proposed framework as a cross-domain method for uncertainty-aware predictive analytics.

IX. DISCUSSION

The findings support the central claim that predictive analytics should not be evaluated only through classification accuracy. In both datasets, models with strong discrimination did not always provide the most reliable probabilities. Logistic regression and random forest produced competitive AUC values, but their calibration errors were substantially higher than those of gradient boosting. If the analysis had relied only on AUC, the study might have identified a different model as best. However, when we included probability reliability, gradient boosting became the strongest decision-support model in both empirical cases.

This result has important methodological implications. AUC measures ranking ability, not probability correctness. A model may rank risky cases above less risky cases while still assigning probabilities that do not match observed event frequencies. For decision support, this distinction matters. A bank, hospital, school, or marketing department does not only need to know which case is more risky than another. It also needs to know whether it can trust a predicted probability as a basis for action. The Brier score, log loss, expected calibration error, and calibration curves therefore provide information that classification metrics cannot.

The probability-band findings are especially important for applied use. The framework separates cases into low-risk, moderate-risk, borderline, high-risk, and very-high-risk categories. In both datasets, the borderline band had observed event rates close to half of the cases. This means that a simple 0.50 classification threshold can hide the most uncertain decisions. Instead of treating all predictions as equally reliable, decision-makers can use the borderline band to trigger manual review, collect additional data, or delay action. This is a practical contribution because uncertainty becomes operational rather than merely statistical.

The Bayesian credible intervals add another layer of interpretation. Observed event rates in probability bands are sample estimates, and they should not be treated as exact values. For example, the customer churn high-risk band had fewer cases than the low-risk band, so its credible interval was wider. This indicates that high event rates in smaller bands should still be interpreted with uncertainty. Such reporting can prevent overconfidence in apparently strong results, especially when sample sizes differ across risk groups.

The two-case design also strengthens the study. Credit default and customer churn differ in domain, variable structure, and decision context, yet the same uncertainty framework produced interpretable results in both cases. This supports the proposed approach's transferability. However, the results also show that transferability does not mean merging unrelated datasets. The correct approach is cross-domain validation: apply the same framework separately to different datasets, then compare the patterns. This avoids methodological confusion while still demonstrating general usefulness.

From a decision-support perspective, the framework encourages a more responsible use of predictive analytics. Instead of reporting a single performance table and selecting the highest-accuracy model, analysts can report a complete uncertainty profile. This profile includes discrimination, probability scoring, calibration, interval uncertainty, and decision bands. This reporting makes model output more understandable for decision-makers who must balance risk, cost, and intervention priorities.

The results also suggest that uncertainty evaluation can improve trust in predictive systems. Trust should not rest on claims that a model is accurate, but on evidence that its probabilities match observed outcomes and that it identifies uncertain cases transparently. In this sense, uncertainty evaluation is not only a technical improvement. It is part of responsible predictive-analytics practice.

X. PRACTICAL IMPLICATIONS

For financial analytics, the findings suggest that credit-risk models should report not only a predicted default class but also a calibrated default probability and risk band. Low-risk clients may be processed automatically, high-risk clients may require risk mitigation, and borderline clients may require additional documentation or manual review. This approach can reduce overconfidence and improve fairness in decision support.

For marketing analytics, the framework helps customer-retention teams distinguish between low-risk customers, high-risk churn customers, and uncertain customers. Using probability bands allows more efficient allocation of intervention resources. High-risk customers may receive stronger retention offers, while borderline customers may be monitored or targeted with lower-cost engagement strategies.

For applied predictive analytics more generally, the framework provides a reporting template. Analysts should present accuracy and AUC together with calibration curves, Brier score, log loss, uncertainty intervals, and decision-band results. This makes predictive outputs more transparent and more useful for decision-makers.

XI. LIMITATIONS AND FUTURE RESEARCH

The study has several limitations. First, it uses two public benchmark datasets, which makes the analysis reproducible but limits control over data-generation conditions. Second, the study evaluates only binary classification tasks. Future research can extend the framework to multiclass prediction, regression, survival analysis, and time-series forecasting. Third, the study uses a fixed train-test split with bootstrap uncertainty intervals. Future work may add repeated cross-validation or nested validation to further examine model stability. Fourth, the Bayesian component is applied to decision-band event rates rather than to full posterior modeling of all model parameters. Future research can estimate full Bayesian predictive models and compare their posterior predictive distributions with non-Bayesian ensemble outputs.

Future studies can also test the framework in healthcare, education, industrial maintenance, and public-policy datasets. Another useful extension would be to examine how decision-makers interpret uncertainty bands in practice. Predictive uncertainty is useful only if decision-makers can communicate it clearly and use it responsibly. Therefore, future research should connect probability-based uncertainty metrics with user-facing decision interfaces.

XII. CONCLUSION

This study developed and evaluated a cross-domain, probability-based framework for evaluating uncertainty in predictive analytics. The framework integrated probabilistic classification, calibration diagnostics, scoring rules, bootstrap intervals, Bayesian credible intervals, and decision-band analysis. Empirical results from credit default and customer churn prediction showed that conventional metrics alone are insufficient for evaluating predictive models used in decision support. In both datasets, gradient boosting produced the strongest uncertainty profile because it combined acceptable discrimination with better Brier score, log loss, and calibration error. The results also showed that probability bands can identify low-risk, moderate-risk, borderline, and high-risk cases in a way that is directly useful for decision-making. Borderline cases had observed event rates close to 50% in both datasets, confirming that uncertainty-aware evaluation can reveal cases where automated decisions should be treated cautiously. The study concludes that trustworthy predictive analytics requires evaluating accuracy, probability reliability, calibration quality, and uncertainty communication together. Predictive models should not only predict outcomes; they should also show how confident and reliable those predictions are.

Author Contributions

The author conducted the conceptualization, methodology, data analysis, investigation, writing, review, editing, and final approval of the manuscript.

Funding

This research received no external funding.

Data Availability

The dataset will be available from the author upon reasonable request.

Conflicts of Interest

The author declares no conflict of interest.

REFERENCES

1. Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv. DOI: 10.48550/arXiv.2107.07511.
2. Ashukha, A., Lyzhov, A., Molchanov, D., & Vetrov, D. (2020). Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. International Conference on Learning Representations.
3. Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. Journal of the American Statistical Association, 112(518), 859-877. DOI: 10.1080/01621459.2017.1285773.
4. Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. Journal of Statistical Software, 76(1), 1-32. DOI: 10.18637/jss.v076.i01.
5. Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis: The TRIPOD statement. Annals of Internal Medicine, 162(1), 55-63. DOI: 10.7326/M14-0697.
6. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. Proceedings of the 33rd International Conference on Machine Learning, 48, 1050-1059.
7. Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian data analysis (3rd ed.). CRC Press. DOI: 10.1201/b16018.

8. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning, 70, 1321-1330.
9. Hullermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Machine Learning, 110, 457-506. DOI: 10.1007/s10994-021-05946-3.
10. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? Advances in Neural Information Processing Systems, 30, 5574-5584.
11. Kruschke, J. K. (2014). Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan (2nd ed.). Academic Press.
12. Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer. DOI: 10.1007/978-1-4614-6849-3.
13. Kuleshov, V., Fenner, N., & Ermon, S. (2018). Accurate uncertainties for deep learning using calibrated regression. Proceedings of the 35th International Conference on Machine Learning, 80, 2796-2804.
14. Kumar, A., Liang, P. S., & Ma, T. (2018). Verified uncertainty calibration. Advances in Neural Information Processing Systems, 31, 3792-3803.
15. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems, 30, 6402-6413.
16. Leibig, C., Allken, V., Ayhan, M. S., Berens, P., & Wahl, S. (2017). Leveraging uncertainty information from deep neural networks for disease detection. Scientific Reports, 7, 17816. DOI: 10.1038/s41598-017-17876-z.
17. Mehrtash, A., Wells, W. M., Tempany, C. M., Abolmaesumi, P., & Kapur, T. (2020). Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE Transactions on Medical Imaging, 39(12), 3868-3878. DOI: 10.1109/TMI.2020.3006437.
18. Moons, K. G. M., Altman, D. G., Reitsma, J. B., Ioannidis, J. P. A., Macaskill, P., Steyerberg, E. W., Vickers, A. J., Ransohoff, D. F., & Collins, G. S. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis: Explanation and elaboration. Annals of Internal Medicine, 162(1), W1-W73. DOI: 10.7326/M14-0698.
19. Naeini, M. P., Cooper, G. F., & Hauskrecht, M. (2015). Obtaining well-calibrated probabilities using Bayesian binning. Proceedings of the 29th AAAI Conference, 2901-2907. DOI: 10.1609/aaai.v29i1.9602.
20. Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., & Tran, D. (2019). Measuring calibration in deep learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 38-41.
21. Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., & Snoek, J. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. Advances in Neural Information Processing Systems, 32, 13991-14002.
22. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144. DOI: 10.1145/2939672.2939778.
23. Riley, R. D., Snell, K. I. E., Ensor, J., Burke, D. L., Harrell, F. E., Moons, K. G. M., & Collins, G. S. (2019). Minimum sample size for developing a multivariable prediction model: Part II - binary and time-to-event outcomes. Statistics in Medicine, 38(7), 1276-1296. DOI: 10.1002/sim.7992.
24. Senge, R., Bosner, S., Dembczynski, K., Haasenritter, J., Hirsch, O., Donner-Banzhoff, N., & Hullermeier, E. (2014). Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. Information Sciences, 255, 16-29. DOI: 10.1016/j.ins.2013.07.030.
25. Thulasidasan, S., Chennupati, G., Bilmes, J., Bhattacharya, T., & Michalak, S. (2019). On mixup training: Improved calibration and predictive uncertainty for deep neural networks. Advances in Neural Information Processing Systems, 32, 13888-13899.
26. Vaicenavicius, J., Widmann, D., Andersson, C., Lindsten, F., Roll, J., & Schon, T. B. (2019). Evaluating model calibration in classification. Proceedings of Machine Learning Research, 89, 3459-3467.
27. Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. BMC Medicine, 17, 230. DOI: 10.1186/s12916-019-1466-7.
28. Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and Computing, 27(5), 1413-1432. DOI: 10.1007/s11222-016-9696-4.
29. Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., Bonten, M. M. J., Damen, J. A. A. G., Debray, T. P. A., De Vos, M., Dhiman, P., Haller, M. C., Harhay, M. O., Henckaerts, L., Kreuzberger, N., Lohmann, A., Luijken, K., Ma, J., Martin, G. P., … van Smeden, M. (2020). Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. BMJ, 369, m1328. DOI: 10.1136/bmj.m1328.
30. Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using stacking to average Bayesian predictive distributions. Bayesian Analysis, 13(3), 917-1007. DOI: 10.1214/17-BA1091.