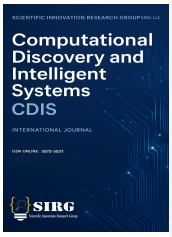

Computational Discovery and Intelligent Systems
CDIS
ISSN: 3070-5037/© 2026 CDIS. (CC BY 4.0) License.
Journal Homepage
https://pub.scientificirg.com/index.php/CDIS

Trustworthy and Explainable Artificial Intelligence for Reliable Bank Term Deposit Subscription Prediction

Gehad Ashraf Abd Elmoneam [a,1], **Heba-t-allah Khalid** [a], **Manar Taha** [a], **and Habiba Salah** [a]

[a] *Department of Computer Science, Faculty of Computers and Information, Suez University, Suez 43221, Egypt;*

E-mails: ggehad894@gmail.com, hebakhalid108@gmail.com, mt9707182@gmail.com, and habibasalahsaad@gmail.com

ABSTRACT

Predicting subscription to a bank term deposit is a highly imbalanced binary classification problem in which apparently strong performance can result from information that is unavailable at the time of targeting. This study develops and audits a leakage aware prediction pipeline using the UCI Bank Marketing dataset. The analysis compares logistic regression with XGBoost, LightGBM, and CatBoost, evaluates class imbalance treatments, performs hyperparameter optimization, examines probability calibration, and adds model agnostic and model specific explanations. The post call duration variable is excluded from the deployable feature set and retained only for a controlled leakage experiment. Five fold cross validation on the training partition gives realistic ROC AUC values from 0.7519 to 0.7959, whereas inclusion of duration increases the corresponding values to 0.8966 to 0.9375. On the untouched random test partition, the calibrated LightGBM model obtains ROC AUC 0.8101, average precision 0.4698, and Brier score 0.0795. An inferred chronological stress test produces ROC AUC 0.7089, demonstrating that random partition estimates can overstate temporal portability. Under illustrative unit economics, a threshold selected from training set out of fold predictions increases test set expected profit from 16,595 to 41,825 units. These findings support a practical principle: leakage control, temporal validation, calibration, and decision analysis are more consequential than small differences among strong tree boosting algorithms.

PAPER INFORMATION

HISTORY

Received: 9 May 2026
Revised: 16 June 2026
Accepted: 28 July 2026
Online: 29 August 2026

MSC

68T07; 68R10; 94A60; 68M15

KEYWORDS

Trustworthy; Explainable Artificial Intelligence; Bank Term Deposit; Subscription Prediction; UCI Bank Marketing.

1 INTRODUCTION

Direct marketing systems must rank prospective customers before a contact is made. The operational objective is therefore not merely to classify historical outcomes, but to estimate which available customers are suitable for a future campaign. The UCI Bank Marketing dataset records telephone campaigns conducted by a Portuguese banking institution and defines a binary target indicating subscription to a term deposit [1]. The dataset is widely used because it contains realistic mixed type predictors, repeated campaign contacts, substantial class imbalance, and a variable that is observed after the call.

The post call duration variable creates a central validity problem. A call duration is not available when a customer is selected for contact, and its value is closely tied to whether the contact reached a meaningful conversation. The UCI documentation explicitly recommends discarding this variable when the objective is realistic prediction [2]. A model that includes duration can therefore provide a high benchmark score while failing the information timing required by deployment.

[1]Corresponding author at Department of Computer Science, Faculty of Computers and Information, Suez University, Suez 43221, Egypt; E-mail: ggehad894@gmail.com

This paper presents a reproducible empirical framework with four aims. First, the effect of duration is quantified rather than described informally. Second, preprocessing and resampling are confined to training folds. Third, performance is assessed with metrics appropriate for an imbalanced target and with an inferred out of time stress test. Fourth, probability calibration, cost sensitive threshold selection, and post hoc explanations are connected to the operational decision.

The key contributions are methodological rather than claims of a universally superior classifier. The study establishes a leakage free baseline, compares three modern gradient boosting systems, evaluates calibration and class imbalance handling, audits the stability of the selected model, and documents where the available evidence does not support stronger claims. The contribution is therefore an auditable experimental design that connects information timing, predictive discrimination, probability quality, and decision utility on one public banking task.

The remainder of the paper is organized as follows. Section 2 reviews bank direct marketing prediction and the methodological literature on heterogeneous tabular models, class imbalance, calibration, explanation, and fairness auditing. Section 3 defines the data, target, and information availability boundary. Section 4 describes feature engineering, model construction, validation, optimization, and the reproducible workflow. Section 5 defines the evaluation metrics and decision utility. Section 6 reports leakage, ablation, held out, temporal, calibration, and explainability results. Section 7 defines the evaluation metrics and the illustrative decision rule. Section 8 reports the leakage, ablation, held out, temporal, calibration, explainability, extension, and fairness results. Section 9 discusses the operational interpretation of these findings. Section 10 states the limitations and reproducibility requirements, and Section 11 concludes.

2 RELATED WORK

2.1 Bank Direct Marketing Prediction

The foundational study by Moro, Cortez, and Rita examined a Portuguese retail bank telemarketing problem using data collected from 2008 to 2013. The study analyzed a substantially larger attribute space, selected a reduced set of predictors, compared logistic regression, decision trees, neural networks, and support vector machines, and used a rolling window evaluation in which more recent observations served as the evaluation period [1]. The reported results showed that neural networks achieved an AUC close to 0.80 and that selecting the upper half of ranked customers could identify a large share of subscribers. The methodological importance for the present work is the explicit recognition that temporal ordering matters and that variables observed during a call are unsuitable for a pre contact targeting system.

Marinakos and Daskalaki focused directly on class imbalance in bank direct marketing. Their Journal of Marketing Analytics study compared statistical, distance based, induction, and machine learning classifiers and evaluated several resampling strategies, including cluster based approaches, with the objective of improving recognition of the underrepresented depositor class [3]. This work motivates the present ablation of class weighting and synthetic sampling. The present study differs in two respects. Resampling is confined to the training fold, and model selection emphasizes average precision and F1 rather than accuracy alone.

Tékouabou et al. proposed a class membership based classifier for heterogeneous bank telemarketing data and emphasized transparent treatment of nominal variables [4]. Their reported accuracy and AUC are substantially higher than the realistic results in the present paper, but the values are not directly comparable because performance depends on the dataset version, feature construction, split design, and availability of call related information. Such differences reinforce the need for explicit dataset provenance and leakage audits rather than direct ranking of isolated headline scores.

Earlier direct marketing research also shows that preprocessing can materially affect classifier sensitivity. Crone, Lessmann, and Stahlbock studied preprocessing effects in direct marketing classification and reported that classifier performance is sensitive to preparation choices [5]. Duman, Ekinci, and Tanriverdi compared classifiers for database marketing with imbalanced datasets and emphasized that the preferred method depends on both the classifier and the class distribution [6]. Together, these studies support an experimental design in which encoding, imbalance treatment, and model family are evaluated separately rather than assumed to have a universal ordering.

2.2 Models for Heterogeneous Tabular Data

XGBoost provides a regularized and scalable tree boosting system that combines additive decision trees with second order optimization and explicit control of model complexity [7]. LightGBM uses histogram based learning and leaf wise tree growth to reduce computational cost while retaining competitive predictive quality [8]. CatBoost introduces ordered target statistics and an ordered boosting procedure designed to reduce prediction shift when categorical variables are used [9]. These methods are related, but they are not interchangeable implementations of a single algorithm. Their different treatment of tree growth, categorical information, and regularization justifies a controlled comparison.

The FT Transformer adapts the Transformer architecture to tabular data by representing features as tokens and applying self attention across feature tokens [10]. Deep tabular models can be competitive, but their evaluation requires careful tuning and sufficiently repeated validation. In the present work the FT Transformer is an exploratory extension because it uses a single stratified validation split for early stopping, whereas the primary tree models use repeated and cross validated procedures.

2.3 Evaluation, Imbalance, Calibration, and Explanation

Saito and Rehmsmeier showed that precision recall analysis can be more informative than ROC analysis when positives are rare because precision directly reflects the fraction of positive predictions that are correct [11]. This principle is central here because the subscription prevalence is approximately 11.7 percent. ROC AUC remains useful as a threshold independent ranking measure, but it is not used as the sole criterion for selecting imbalance treatment or the deployable model.

SMOTE creates synthetic minority observations by interpolating between minority examples and remains a standard comparator for imbalanced classification [12]. Its use before cross validation can leak information from validation observations into the synthetic training set. The implementation therefore places each sampler inside the fold specific training pipeline [13]. Probability calibration is treated as a separate property from discrimination. Niculescu Mizil and Caruana demonstrated that different classifiers can have similar discrimination but materially different probability quality [14]. The present analysis uses isotonic calibration and reports Brier score before and after calibration.

SHAP provides a unified additive attribution framework that assigns feature contributions to individual predictions and aggregates those contributions for global summaries [15]. The present paper uses SHAP to describe the fitted model, not to infer causal effects or claim that a customer characteristic should be manipulated. Local waterfall plots are interpreted as decompositions of a model output relative to a background expectation. They are not treated as causal explanations, individual treatment recommendations, or evidence that a feature can be changed without changing other variables.

The literature also supports separating model search diagnostics from final predictive evidence. The Optuna curves in this paper show trial scores and the best score observed so far on the training search subsample. They document the behavior of the search procedure, not an independent estimate of generalization. This distinction is consistent with the broader methodological concern that model comparisons depend on the number of methods, partitions, and selection decisions considered [16]. The present paper therefore reports the search traces as configuration evidence and reserves the held out test partition for the primary final comparison.

Table 1 positions the present study against the most relevant verified papers. Reported scores are intentionally not combined into a single ranking because the cited studies use different dataset versions, feature sets, targets, and validation protocols.

Table 1: Verified related work and methodological position of the present study

Study	Data and task	Methods	Evaluation emphasis	Relation to present study
Moro et al., 2014 [1]	Portuguese bank telemarketing; deposit subscription	Logistic regression, decision tree, neural network, support vector machine	Rolling window, AUC and lift	Provides the principal precedent for temporal evaluation and operational ranking.
Duman et al., 2012 [6]	Imbalanced database marketing	Alternative statistical and machine learning classifiers	Class imbalance and classifier comparison	Supports treating classifier and prevalence as interacting design factors.
Crone et al., 2006 [5]	Direct marketing classification	Preprocessing and classifier sensitivity analysis	Effect of preparation choices on predictive sensitivity	Motivates the controlled encoding and preprocessing ablations.
Marinakos and Daskalaki, 2017 [3]	Public bank direct marketing data	Multiple classifiers and resampling strategies	Recognition of the rare depositor class	Motivates fold confined resampling and prevalence aware metrics.
Tékouabou et al., 2022 [4]	Bank telemarketing prediction with heterogeneous variables	Class membership based classifier	Transparency and nominal variable treatment	Shows the value of transparent categorical handling; scores are not directly comparable across dataset and split designs.

3 DATA AND PROBLEM DEFINITION

The analysis uses the full UCI file `bank-full.csv`, containing 45,211 records, 16 input variables, and one binary target. The repository identifies the source as direct marketing by a Portuguese banking institution and documents no missing values in the released file [17]. The target is defined as

$$y_i = \begin{cases} 1, & \text{if customer } i \text{ subscribed,} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The supplied notebook reports an 80 percent stratified training partition with 36,168 records and a 20 percent test partition with 9,043 records. The test prevalence is 0.1170. The input variables include age, job, marital status, education, credit default status, account balance, housing loan status, personal loan status, contact type, calendar information, campaign history, and previous campaign outcome.

Table 2 summarizes the principal data quantities used in the manuscript. **Figure 1** presents selected empirical subscription rates. These descriptive rates are not causal effects and do not by themselves imply that contacting a subgroup will increase subscription.

Table 2: Dataset and partition summary

Quantity	Value
Total records	45,211
Input variables	16
Training records	36,168
Held out test records	9,043
Test subscription rate	0.1170
Realistic numeric predictors	13
Categorical predictors	9

4 LEAKAGE AWARE FEATURE ENGINEERING

All row wise transformations are deterministic functions of the current record. The realistic representation includes raw age and day, cyclical encodings of month and day, a signed logarithmic account balance, a negative balance indicator, logarithmic campaign counts, a previous contact indicator, a previous success indicator, and an age group. For a nonnegative count z, the transformed value is

$$z_{\log} = \log(1 + z). \tag{2}$$

For a signed balance b, the transformation is

$$b_{\text{slog}} = \operatorname{sign}(b) \log(1 + |b|). \tag{3}$$

Cyclical month and day representations use

$$s(v) = \sin\left(\frac{2\pi v}{P}\right), \qquad c(v) = \cos\left(\frac{2\pi v}{P}\right), \tag{4}$$

where P is the corresponding calendar period. The sentinel value -1 in `pdays` is mapped to zero for the logarithmic representation and is separately represented by a contacted indicator.

Two feature scenarios are retained. The realistic scenario excludes duration. The full scenario adds $\log(1 + \texttt{duration})$ and is used only to quantify optimistic bias. **Figure 2** shows the strong separation of the duration distribution by outcome. This visual separation is consistent with the information timing concern, but it is not evidence that duration is a valid pre contact predictor.

5 EXPERIMENTAL PROTOCOL

The held out test set is separated before fitting encoders, scalers, target encoders, or resamplers. A column transformer standardizes numeric variables and one hot encodes categorical variables. The selected encoding is one hot encoding

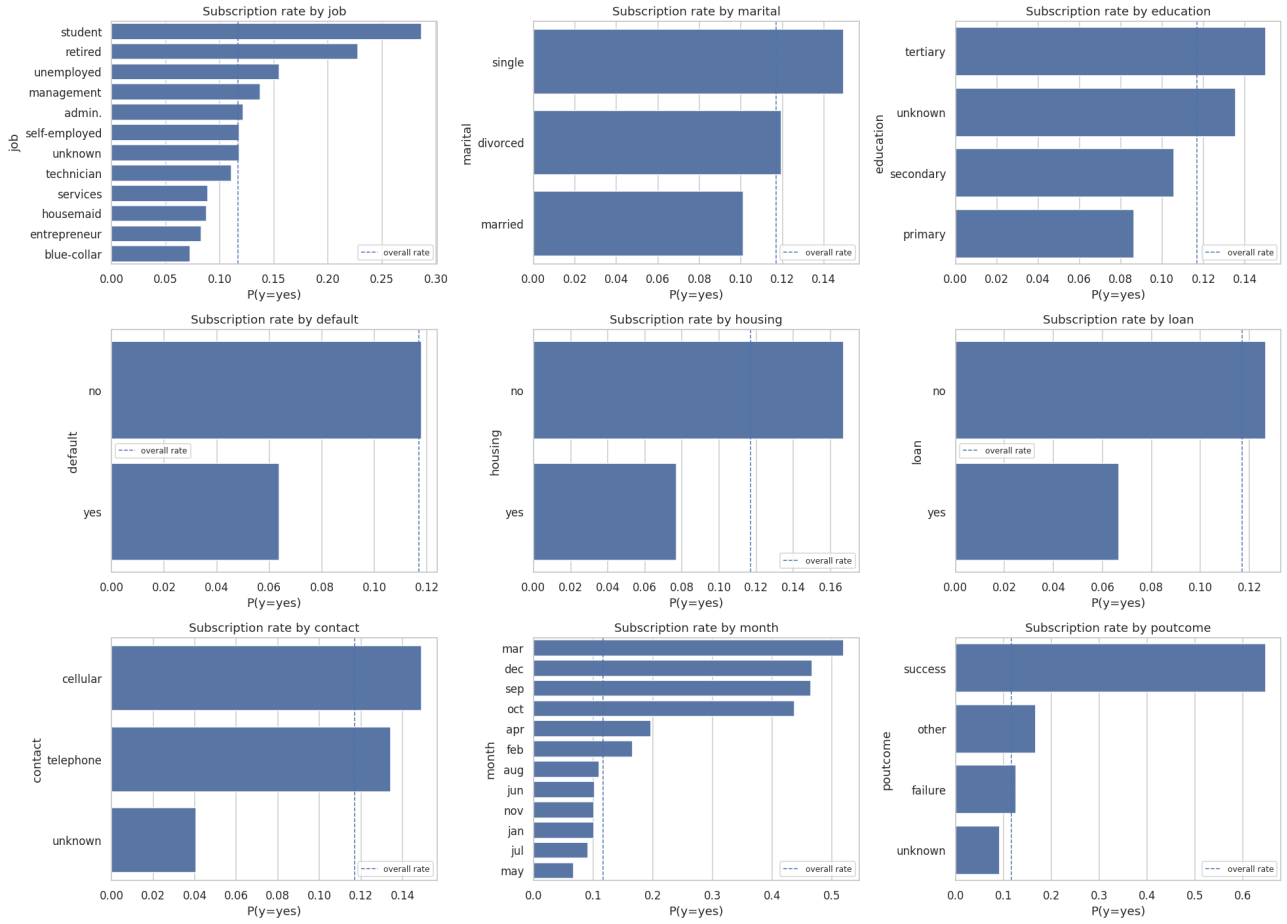

Figure 1: Empirical rates by selected demographic, campaign, and contact variables. The dashed line denotes the overall rate

because target encoding improves training set cross validated ROC AUC from 0.7915 to 0.7929, a gain below the prespecified 0.002 switching threshold, while one hot encoding retains simpler category level interpretation.

Five fold stratified cross validation is used for the main training comparisons. The class imbalance ablation compares no correction, inverse frequency class weighting, SMOTE, Borderline SMOTE, and SMOTE with Tomek link cleaning. Resampling is executed inside an imbalanced learning pipeline, so synthetic samples are created only within the relevant training fold. Strategy selection uses F1 as the primary criterion and ROC AUC as a tie breaker within 0.005 F1 of the best strategy.

XGBoost, LightGBM, and CatBoost are tuned with 30, 30, and 25 Optuna trials, respectively, on a fixed stratified subsample of 12,000 training records. The objective is average precision. Each search jointly selects model hyperparameters and either a class weight multiplier or a resampling method. The selected configuration is refit on all training records. The test set is not used during this search.

The primary candidate set contains logistic regression, tuned XGBoost, tuned LightGBM, tuned CatBoost, a soft voting ensemble, and a stacking ensemble. The ensemble estimators use the tuned tree parameters and class weights. The calibrated deployment candidate is selected using training set five fold average precision, after which isotonic calibration is fit with cross fitting. The test set is then used for the principal candidate table.

5.1 Model Overview

The study uses a deliberately heterogeneous model set. Logistic regression provides a transparent linear reference and follows the binary response formulation introduced by Cox [18]. XGBoost, LightGBM, and CatBoost represent three distinct gradient boosting implementations. XGBoost emphasizes regularized additive tree boosting [7]; LightGBM uses histogram based learning and leaf wise growth [8]; and CatBoost uses ordered procedures designed for categorical features [9]. The three models are tuned under the same fold confined data preparation and are compared with identical primary metrics.

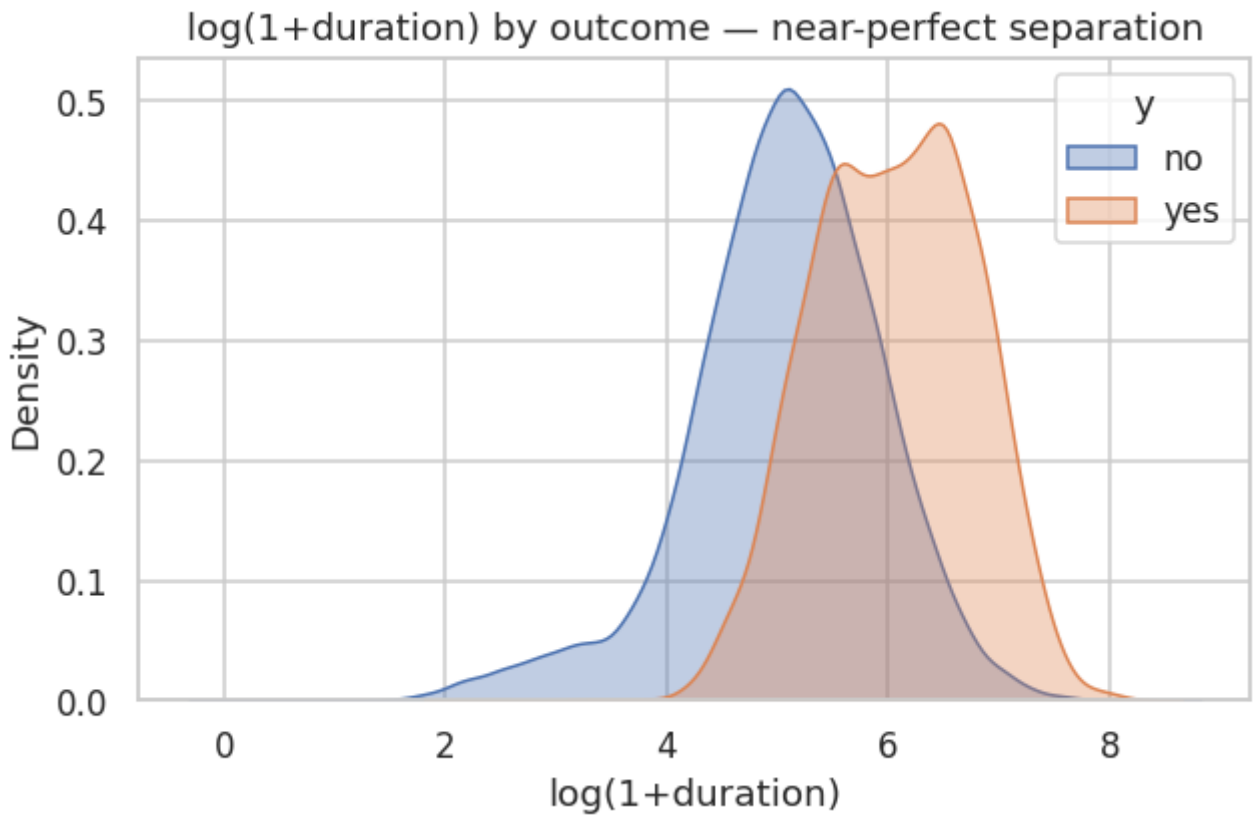

Figure 2: Distribution of logarithmic call duration by outcome

The soft voting ensemble is a probability averaging rule. It combines the outputs of the three tuned tree models without adding a learned meta model. The stacking ensemble follows stacked generalization [19]: out of fold base predictions are supplied to a logistic regression meta learner, which learns how to combine the base model outputs. In this implementation the ensemble wrappers use class weighting only, while the individual tree pipelines retain the imbalance treatment selected for each model.

The FT Transformer is an exploratory attention based model that tokenizes numerical and categorical features and applies self attention across tokens, following the tabular Transformer formulation of Gorishniy et al. [10]. Its separate validation based early stopping procedure prevents a fully symmetric comparison with the primary tree model search. A cost sensitive LightGBM extension changes the training objective through explicit sample weights. This extension is motivated by cost sensitive classification, in which different error types have different penalties [20].

Calibration and explanation are analysis components rather than additional predictive model families. Isotonic regression is applied after candidate selection to improve probability quality, following the distinction between discrimination and probability calibration discussed by Niculescu Mizil and Caruana [14]. SHAP is used for global and local model attribution according to the additive explanation framework of Lundberg and Lee [15]. For a transformed input vector x_i, its logistic probability estimate is given in **Equation 5**, the standard logistic link formulation.

$$p_i = \sigma(\beta_0 + x_i^{\top}\beta), \qquad \sigma(a) = \frac{1}{1 + \exp(-a)}. \tag{5}$$

The model is fitted with balanced class weights and provides a useful lower complexity comparison.

The three tree models learn additive functions of decision trees. The additive representation in **Equation6** is the common boosting form used to describe the three implementations, while their optimization and categorical handling differ [7, 8, 9].

$$f(x_i) = \sum_{m=1}^{M} f_m(x_i), \qquad p_i = \sigma(f(x_i)), \tag{6}$$

where M is the number of boosting iterations. XGBoost uses regularized second order boosting, LightGBM uses histogram based gradient boosting with leaf wise growth, and CatBoost uses ordered treatment of categorical information. The manuscript does not assume that any one implementation is universally best. Each is tuned with the same training data discipline and evaluated using the same metrics.

The soft voting model averages the predicted probabilities of the tree learners as shown in **Equation 7**. This is a probability aggregation rule used by the implemented ensemble, not an additional fitted classifier family.

$$p_{\text{vote}}(x) = \frac{1}{K} \sum_{k=1}^{K} p_k(x). \tag{7}$$

The stacking model supplies base learner probabilities to a logistic regression meta learner. Internal cross fitting produces the meta learner training inputs, which reduces the risk that a base learner's in sample predictions are mistaken for out of sample evidence.

The FT Transformer extension first maps each numerical feature to a learned token and each categorical feature to an embedding token. A Transformer encoder applies self attention across the feature tokens. The classification head uses the output corresponding to a learned classification token. This architecture is included to test a distinct inductive bias, but the single validation split used for early stopping prevents a fully comparable claim against the primary tree models.

5.2 End to End Pseudocode

Algorithm 1 summarizes the implemented workflow. The pseudocode separates the realistic and full feature scenarios, keeps all learned preprocessing inside the relevant training operation, and prevents the held out test partition from entering primary model or threshold selection. The symbol `SEED` denotes the notebook seed variable. Because the exported audit does not expose its numeric value, the pseudocode does not invent one.

The pseudocode is an implementation summary rather than a claim that every notebook cell is represented line by line. Four reconciliation points are important for faithful interpretation. SMOTE plus Tomek link cleaning belongs to the fixed imbalance ablation, whereas the joint Optuna search uses ADASYN as its third resampling option. Random Forest is not part of the final model zoo or ensemble after the model reduction pass. The fairness routines generate exploratory audit tables, but fairness results are not interpreted as primary findings unless their numerical outputs are archived with the release. Finally, the FT Transformer, passthrough stacking, enhanced feature, and training time cost sensitive variants are extensions and do not redefine the primary candidate or threshold conclusions.

5.3 Hyperparameter Search and Selection Logic

The Optuna search uses average precision as its objective because the minority class is the operational focus. The search space includes tree count, depth or leaf count, learning rate, subsampling, column subsampling, minimum child size, regularization, and one imbalance treatment. A treatment trial either selects a class weight multiplier or a sampler with a specified post sampling ratio and neighborhood size. The search uses a stratified 12,000 record subset only for computational efficiency. The final selected parameters are refit on all 36,168 training records.

Selection is deliberately separated into three decisions. First, the encoding and imbalance strategy are chosen using cross validation on the training partition. Second, the calibrated single model is chosen by training partition average precision. Third, the frozen model and thresholds are evaluated on the held out test partition. The later FT Transformer and cost sensitive extensions are marked exploratory because they reuse the test partition after the main comparison.

6 MODEL CONFIGURATION

This section records the configuration used for the reported experiments so that the comparison is reproducible and the distinction between tuned and fixed settings is explicit. The main feature scenario excludes duration. Numeric features are standardized with training fold statistics, and categorical features are one hot encoded with unknown categories ignored. The target encoding ablation is not used in the final model because its ROC AUC gain over one hot encoding is below the prespecified switching threshold.

Table 3 summarizes the shared configuration. The number of parallel workers is set to minus one for individual tree models and to one for the ensemble wrapper to avoid nested parallel execution.

Table 4 gives the exact best configurations recorded by the Optuna searches. The imbalance fields are part of the search and are not ordinary classifier parameters. For XGBoost, the selected SMOTE sampler is rebuilt for the final fit with sampling strategy 0.3491 and nine nearest neighbors. For LightGBM and CatBoost, the selected weight only treatments use multipliers of 0.4560 and 0.5997, respectively, applied to the training negative to positive ratio. The base ratio is computed from the training labels.

Algorithm 1 Leakage safe bank term deposit prediction pipeline

Require: Raw file `bank-full.csv`, seed variable `SEED`, feature specification
Ensure: Fitted candidates, held out metrics, explanations, and audit records

Load the semicolon separated file and set $y_i \leftarrow \mathbb{1}[\texttt{y}_i = \texttt{yes}]$.
$(\mathcal{D}_{\text{tr}}, \mathcal{D}_{\text{te}}) \leftarrow \text{StratifiedSplit}(\mathcal{D}, 0.20, \texttt{SEED})$. ▷ The test set is reserved for primary final evaluation.
Apply the deterministic feature function to both partitions.
    Create `pdays_contacted`, logarithmic counts, signed logarithmic balance,
    cyclical month and day variables, previous success, and age group.
    Define $\mathcal{F}_{\text{realistic}}$ by excluding `duration`; define
    $\mathcal{F}_{\text{full}}$ by adding $\log(1 + \texttt{duration})$ for leakage analysis.
for scenario in {realistic, full} **do**
    for model in {LogReg, XGBoost, LightGBM, CatBoost} **do**
        Compute five fold stratified CV ROC AUC, average precision, and F1 on $\mathcal{D}_{\text{tr}}$.
    end for
end for
Report the duration gap and use $\mathcal{F}_{\text{realistic}}$ thereafter.
for encoding in {one hot, frequency, target} **do**
    Compute five fold CV ROC AUC with fixed LightGBM and fold confined preprocessing.
end for
Select one hot encoding unless another encoding exceeds it by more than 0.002 ROC AUC.
for strategy in {none, class weight, SMOTE, Borderline SMOTE, SMOTE plus Tomek} **do**
    Compute fold confined CV ROC AUC, average precision, F1, and balanced accuracy.
end for
Select the recorded imbalance strategy using the prespecified F1 constrained rule.
Draw a stratified training subsample of at most 12,000 records.
for model in {XGBoost, LightGBM, CatBoost} **do**
    Run Optuna TPE for 30, 30, and 25 trials, respectively, maximizing five fold CV average precision.
The imbalance search uses weight only, SMOTE, Borderline SMOTE, or ADASYN.
end for
Refit the best tree configurations on all training records.
Build soft voting and stacking ensembles from the three tuned tree models.
Use scale pos weight in ensemble base learners and no individual resampler.
Select the single calibration candidate by training only five fold CV average precision.
Fit five fold isotonic calibration on the training partition.
Generate training out of fold probabilities and select the F1 and study defined profit thresholds.
Evaluate primary candidates once on the held out partition using ROC AUC, bootstrap intervals,
average precision, thresholded metrics, MCC, Brier score, confusion matrices, and profit.
Run repeated five by two fold robustness and three seed sensitivity analyses.
Generate calibration, SHAP, permutation, local, dependence, segmentation, and exploratory fairness artifacts.
Run the inferred calendar based out of time stress test and label it exploratory.
Run enhanced feature, passthrough stacking, FT Transformer, and cost sensitive extensions separately.
Archive code, parameters, predictions, figures, derived tables, environment metadata, and audit logs.

The ensembles use the tuned tree parameters and class weights, but the recorded implementation does not pass the individual resamplers through the scikit learn voting and stacking wrappers. This distinction is important. The individual XGBoost pipeline uses SMOTE, while the corresponding XGBoost base estimator inside the ensembles uses class weighting only. The soft voting model averages three tree probabilities. The stacking model uses the same three base learners, internal five fold cross fitting, and a class weighted logistic regression meta learner without feature passthrough.

Table 5 records the calibration and exploratory extension settings. The FT Transformer uses 32 dimensional tokens, four attention heads, three encoder layers, GELU activation, dropout 0.1, AdamW learning rate 10^{-3}, weight decay 10^{-5}, batch sizes 512 and 1024, maximum 20 epochs on CPU or 60 epochs on a GPU, and patience eight. Its loss uses the training negative to positive ratio as the positive class weight. The cost sensitive LightGBM uses the tuned LightGBM parameters without additional frequency weighting and assigns weight 75 to positive training rows and weight 5 to negative training rows.

Table 3: Shared model and pipeline configuration

Component	Configuration
Feature scenario	Realistic scenario with duration excluded
Numeric processing	StandardScaler fitted within each training fold
Categorical processing	OneHotEncoder with unknown categories ignored
Primary validation	Five fold StratifiedKFold, shuffled, fixed seed
Robustness validation	Repeated StratifiedKFold, five folds and two repeats
Optimization objective	Average precision on a stratified 12,000 record subset
XGBoost trials	30 Optuna trials
LightGBM trials	30 Optuna trials
CatBoost trials	25 Optuna trials
Final refit	All 36,168 training records
Threshold grid	0.01 to 0.99 in increments of 0.01
Calibration	Isotonic regression with five fold cross fitting

Table 4: Recorded Optuna configurations for the three tuned tree models

Setting	XGBoost	LightGBM	CatBoost
Tree or iteration count	$n\_estimators = 444$	$n\_estimators = 347$	$iterations = 232$
Tree complexity	$max\_depth = 4$; $min\_child\_weight = 7$	$num\_leaves = 56$; $min\_child\_samples = 57$	$depth = 5$
Learning rate	0.0282976	0.0146222	0.0561029
Row subsampling	0.7319451	0.8114465	Not used
Column subsampling	0.9307439	0.5617394	Not used
Regularization	$gamma = 0.2808254$; $reg\_alpha = 1.6568303$; $reg\_lambda = 0.4914058$	$reg\_alpha = 0.3910046$; $reg\_lambda = 0.4050253$	$l2\_leaf\_reg = 2.8745027$
Imbalance method	SMOTE	Weight only	Weight only
Imbalance parameters	sampling strategy 0.3491285; $k = 9$	weight multiplier 0.4559595	weight multiplier 0.5996903
Fixed implementation settings	histogram tree method; evaluation metric AUC; random state seed	random state seed; verbosity disabled	random seed; verbosity disabled; file writing disabled
Search PR AUC	0.4127105	0.4208958	0.4242783

Table 5: Calibration, ensemble, and exploratory extension configuration

Component	Configuration
Voting ensemble	Soft probability average of XGBoost, LightGBM, and CatBoost; wrapper workers one
Stacking ensemble	Same three base models; logistic regression meta learner; internal $cv = 5$; no passthrough
Calibrated model	LightGBM selected by training five fold PR AUC; isotonic calibration; $cv = 5$
FT Transformer tokens	One token per numeric or categorical feature plus learned classification token
FT Transformer encoder	Token dimension 32; four heads; three layers; feed forward width 128; GELU; dropout 0.1
FT Transformer training	AdamW; learning rate 10^{-3}; weight decay 10^{-5}; batch sizes 512 and 1024; patience eight
Cost sensitive model	LightGBM with positive weight 75 and negative weight 5; no additional class frequency weighting

7 METRICS AND DECISION RULES

For a score p_i and threshold t, the predicted class is defined in **Equation 8**. This thresholding rule is the decision interface used for the reported F1 and illustrative profit comparisons.

$$\hat{y}_i(t) = \mathbb{1}[p_i \geq t]. \tag{8}$$

The reported classification measures include precision, recall, F1, balanced accuracy, and Matthews correlation coefficient. With confusion matrix entries (TP, TN, FP, FN), the F1 measure is defined in **Equation 9**, a harmonic mean of precision and recall. This confusion matrix formulation is consistent with the evaluation framework discussed by Saito and Rehmsmeier for imbalanced binary classification [11].

$$\mathrm{F1} = \frac{2TP}{2TP + FP + FN}, \tag{9}$$

The Matthews correlation coefficient is defined in **Equation 10**, using all four confusion matrix cells. The same four cell structure is used in the imbalanced classification treatment of Saito and Rehmsmeier [11].

$$\begin{aligned} \mathrm{MCC} &= \frac{TP\,TN - FP\,FN}{\sqrt{D}}, \\ D &= (TP + FP)(TP + FN)(TN + FP)(TN + FN). \end{aligned} \tag{10}$$

Average precision is reported as the area under the precision recall construction used by the implementation. Its use is motivated by the prevalence sensitivity of precision recall analysis under class imbalance [11]. The no skill average precision reference is the test prevalence, 0.1170, and the reported lift is average precision divided by this prevalence.

The Brier score is the mean squared error of probabilistic predictions, as defined in **Equation 11**. Its use here follows the probability score interpretation and decomposition introduced by Murphy [21].

$$\mathrm{Brier} = \frac{1}{n}\sum_{i=1}^{n}(p_i - y_i)^2. \tag{11}$$

For the illustrative decision analysis, contacting a customer costs $C_c = 5$ units and a successful subscription yields net benefit $B_s = 80$ units before contact cost. The study defined expected profit as given in **Equation 12**. This equation is an explicit utility model for the present experiment rather than a universal banking law. The decision theoretic motivation for making error costs explicit is consistent with cost sensitive learning [20].

$$\Pi(t) = TP(t)(B_s - C_c) - FP(t)C_c. \tag{12}$$

These values are not institutional cost estimates. The profit threshold is selected from training set out of fold predictions and then applied to the test scores.

8 RESULTS

8.1 Leakage Quantification

Table 6 reports training set five fold cross validation. Duration increases ROC AUC by 0.1447 for logistic regression, 0.1404 for XGBoost, 0.1435 for LightGBM, and 0.1416 for CatBoost. The effect is therefore not specific to a single implementation. **Figure 3** visualizes this contrast. Subsequent model selection, explanation, and business analysis use the realistic scenario only.

Table 6: Training set five fold cross validation for realistic and full feature scenarios

Model	Realistic	Full	Gap	Full PR AUC
Logistic regression	0.7519	0.8966	0.1447	0.5394
XGBoost	0.7956	0.9360	0.1404	0.6282
LightGBM	0.7915	0.9350	0.1435	0.6264
CatBoost	0.7959	0.9375	0.1416	0.6347

8.2 Encoding and Imbalance Ablations

The encoding ablation produced ROC AUC values of 0.7915 for one hot encoding, 0.7909 for frequency encoding, and 0.7929 for target encoding. The latter did not pass the switching threshold. The imbalance ablation selected class weighting by F1. The class weighted model obtained F1 0.4558 and balanced accuracy 0.7330, compared with F1 0.3598 and balanced accuracy 0.6168 without correction. No correction had the highest ROC AUC, 0.7954, illustrating why ROC AUC alone is insufficient for selecting an operating configuration under rare positives. **Figure 4** shows the ablation.

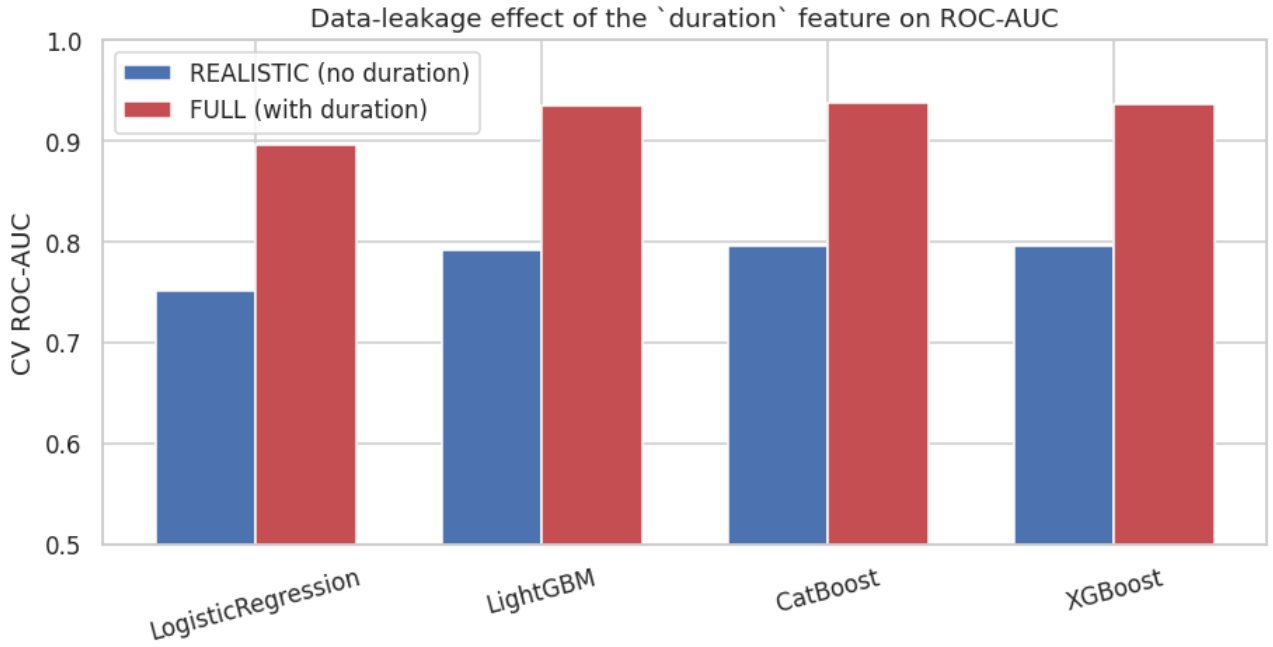

Figure 3: ROC AUC inflation associated with including call duration

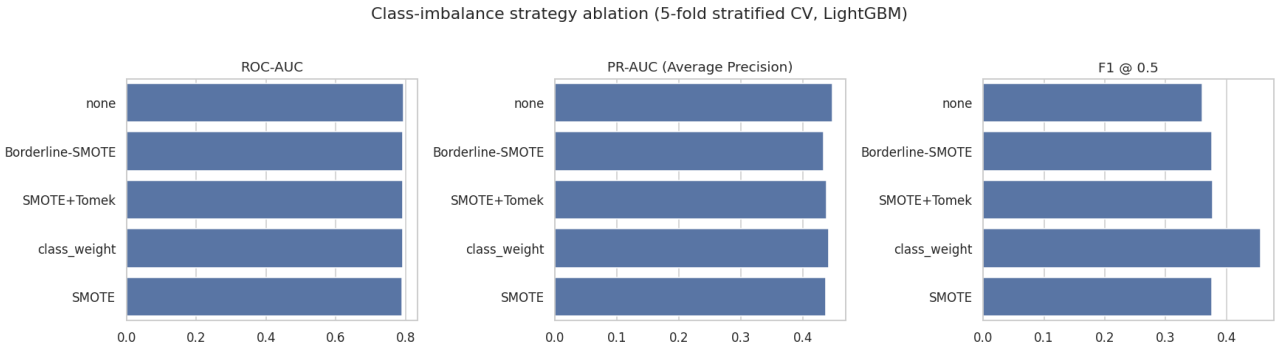

Figure 4: Class imbalance strategy ablation for LightGBM under five fold stratified cross validation

8.3 Held Out Evaluation

Table 7 reports the untouched random test comparison. Stacking has the highest ROC AUC at 0.8102, but the calibrated LightGBM is selected for deployment because the selection rule uses training set cross validated average precision and because calibrated probabilities are required for the subsequent decision analysis. The calibrated model has ROC AUC 0.8101, average precision 0.4698, and Brier score 0.0795. **Figure 5** shows the ROC and precision recall curves, and **Figure 6** shows the confusion matrices at threshold 0.5.

Table 7: Held out test performance at threshold 0.5 unless otherwise stated

Model	ROC AUC	95% low	95% high	PR AUC	Balanced Accuracy	Precision	Recall	F1
Stacking ensemble	**0.8102**	0.7947	0.8257	0.4681	**0.7490**	0.3580	0.6531	0.4625
Voting ensemble	0.8101	0.7946	0.8256	**0.4707**	0.6940	0.5352	0.4386	0.4821
Calibrated LightGBM	0.8101	**0.7949**	**0.8261**	0.4698	0.6010	**0.6638**	0.2164	0.3264
LightGBM	0.8097	0.7941	0.8254	0.4702	0.7194	0.4913	0.5085	**0.4998**
CatBoost	0.8069	0.7913	0.8226	0.4598	0.7263	0.4602	0.5359	0.4952
XGBoost	0.8007	0.7847	0.8164	0.4430	0.6024	0.6411	0.2212	0.3289
Logistic regression	0.7595	0.7431	0.7762	0.3801	0.6930	0.2403	**0.6645**	0.3529

The McNemar comparison between stacking and voting at threshold 0.5 gives a corrected statistic of 347.7554 with $p < 0.001$. This result concerns paired error patterns, not a general proof that the stacking model has superior ranking quality. The difference in ROC AUC between the two models is only 0.0001 and the bootstrap intervals overlap substantially.

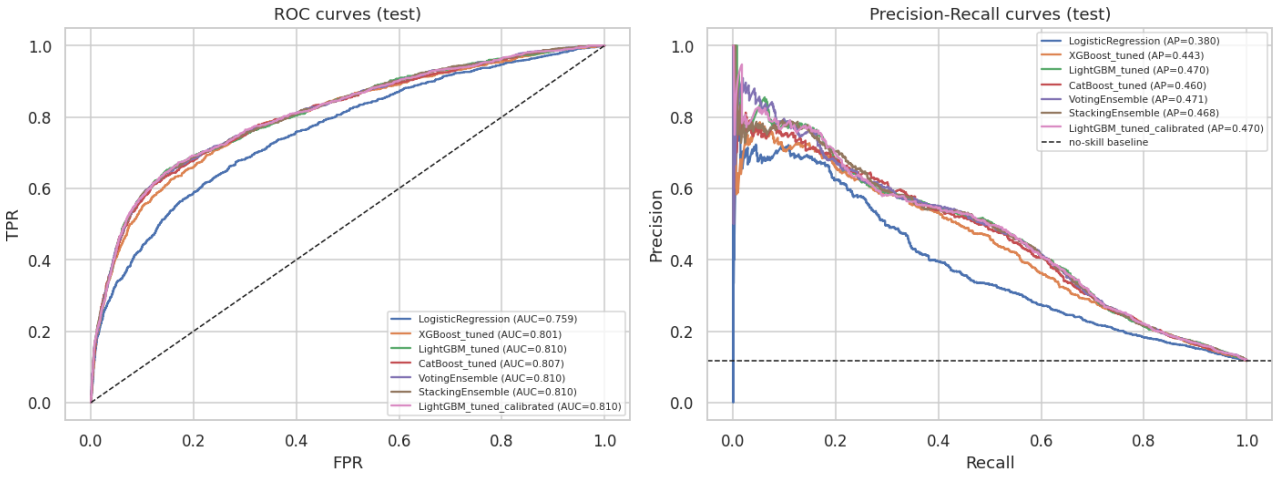

Figure 5: ROC and precision recall curves on the held out test set

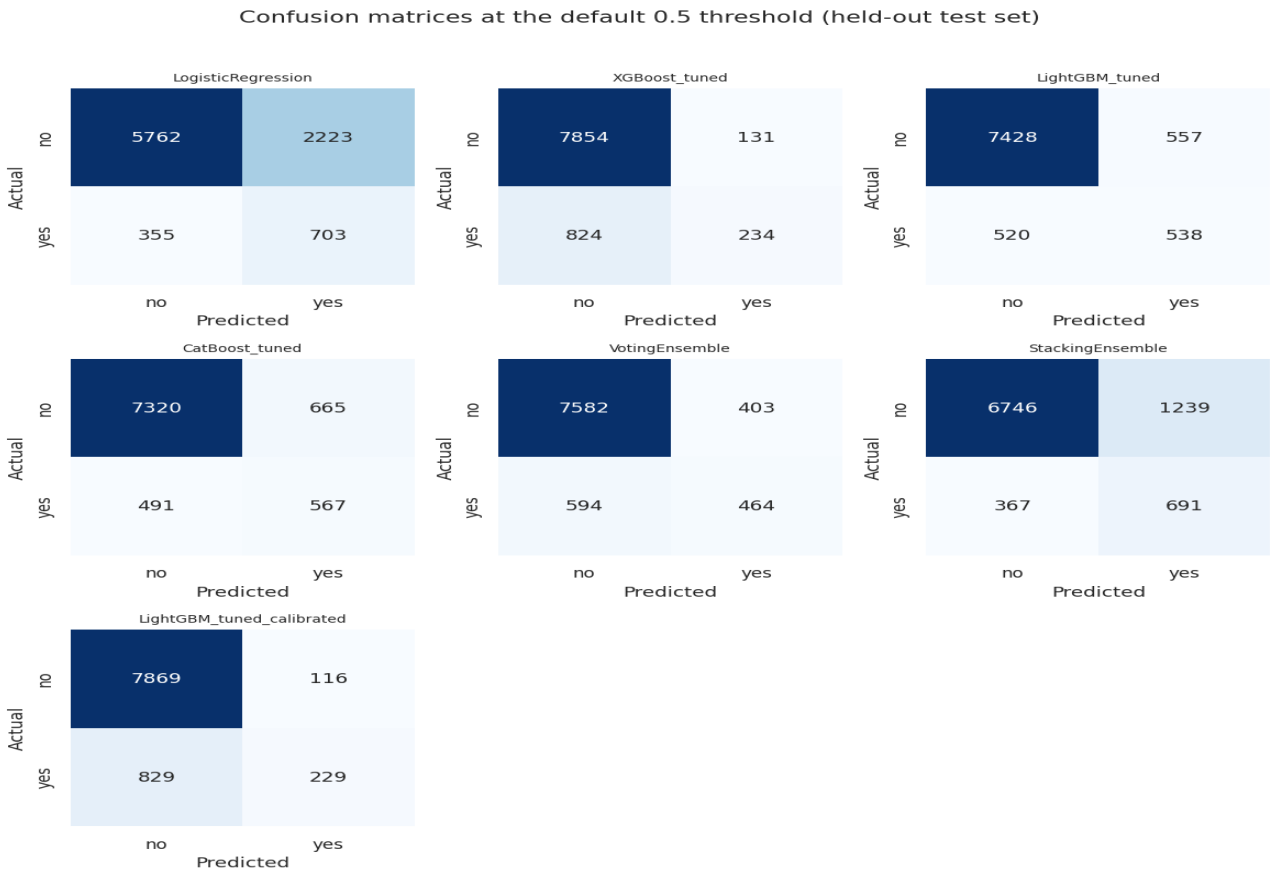

Figure 6: Confusion matrices for the final candidates at threshold 0.5 on the held out test set

8.4 Temporal Stress Test

The released full file is ordered by date, but it does not provide an explicit year field in the version used here. Sorting by the available month index and reserving the final calendar quarter gives 39,710 early records with subscription rate 0.1056 and 5,501 later records with rate 0.1991. The LightGBM stress test obtains ROC AUC 0.7089 and average precision 0.4656, compared with random split five fold ROC AUC 0.7915. This result is an inferred calendar order stress test, not a definitive future year validation. **Figure 7** presents the learning curve of the tuned LightGBM model, showing that cross-validated ROC AUC remains below the training ROC AUC while approaching a plateau as the training sample size increases.

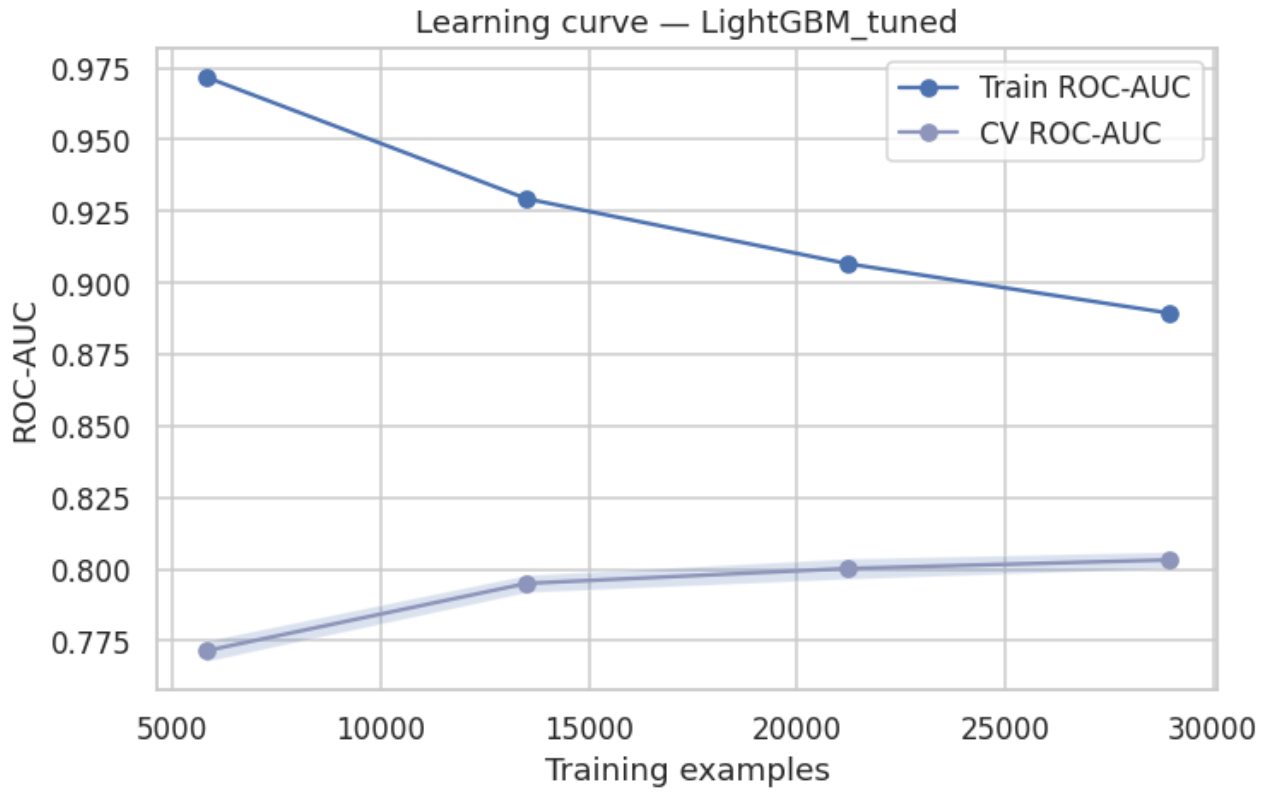

Figure 7: Learning curve for the tuned LightGBM model

8.5 Explanations

The global SHAP ranking identifies contact type, month, contact history, marital status, housing status, prior campaign outcome, age, and transformed financial variables as influential. The permutation ranking gives a similar ordering, with contact, month encodings, day, balance, previous outcome, campaign history, and housing among the leading variables. **Figures 8** Figure 8 shows the SHAP beeswarm, illustrating the direction and magnitude of feature contributions for the tuned LightGBM model. **Figure 9** shows the subscription rates of SHAP derived segments, and **Figures 10** and **11** show two local waterfall explanations. The local plots are descriptive decompositions of fitted outputs and do not establish causal effects.

A SHAP based clustering extension groups explanation vectors rather than raw customer attributes. The supplied output reports segment subscription rates of approximately 0.62, 0.41, 0.08, and 0.03 for the four displayed segments. These groups are exploratory explanation profiles, not validated marketing personas and not causal customer types. The attention based FT Transformer extension obtains test ROC AUC 0.8014, average precision 0.4411, and F1 0.4496. A training time cost weighted LightGBM extension obtains ROC AUC 0.8089, average precision 0.4660, recall 0.7940, F1 0.3455, and profit 48,180 at threshold 0.5. Both extensions reuse the held out test set after the primary comparison and are therefore exploratory rather than confirmatory.

8.6 Subgroup Fairness Audit

The subgroup audit applies the disparate impact ratio and the true positive and false positive rate ranges of Hardt et al. [22] to the frozen LightGBM classifier at the default 0.5 threshold on the held out test partition, grouped by job category, marital status, and age group. **Table 8** reports the results. For job category, the selection rate ratio is 0.087, with a true positive rate range of 0.273 and a false positive rate range of 0.100 across the twelve job categories; the entrepreneur category receives essentially no positive predictions despite a nonzero actual positive rate. For marital status, the selection rate ratio is 0.682, with divorced customers selected at a substantially lower rate than single customers. For age group, the selection rate ratio is 0.652, with older and younger customers selected at higher rates than middle aged customers.

All three selection rate ratios fall below the conventional four fifths rule threshold of 0.80 [23], and the true positive and false positive rate ranges confirm that the disparity is not confined to raw selection rates alone. This finding does not imply that the input variables cause the observed differences; job category, marital status, and age are correlated with the engineered features, including account balance and contact history, so a mitigation strategy would need to address these

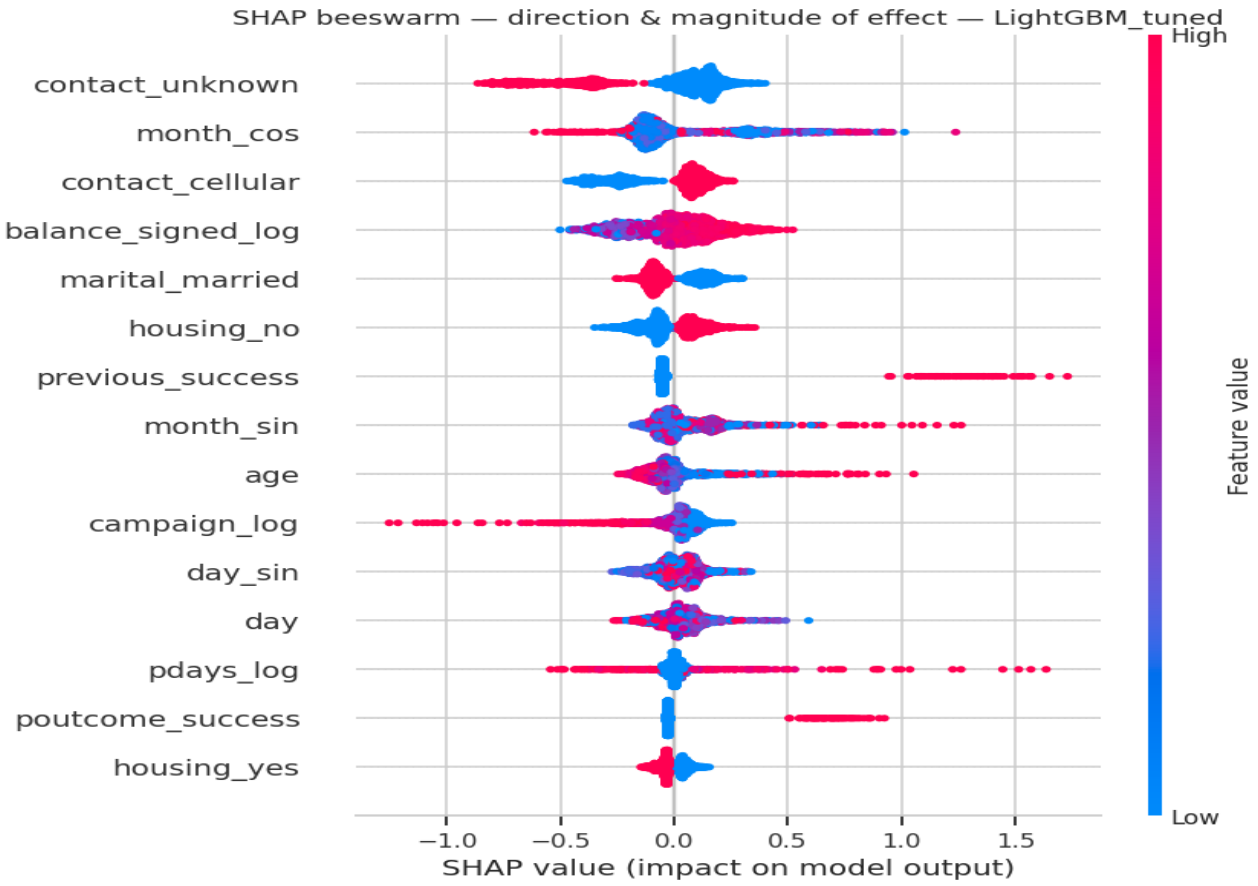

Figure 8: Supplied LightGBM SHAP beeswarm showing feature value, direction, and magnitude of local attribution

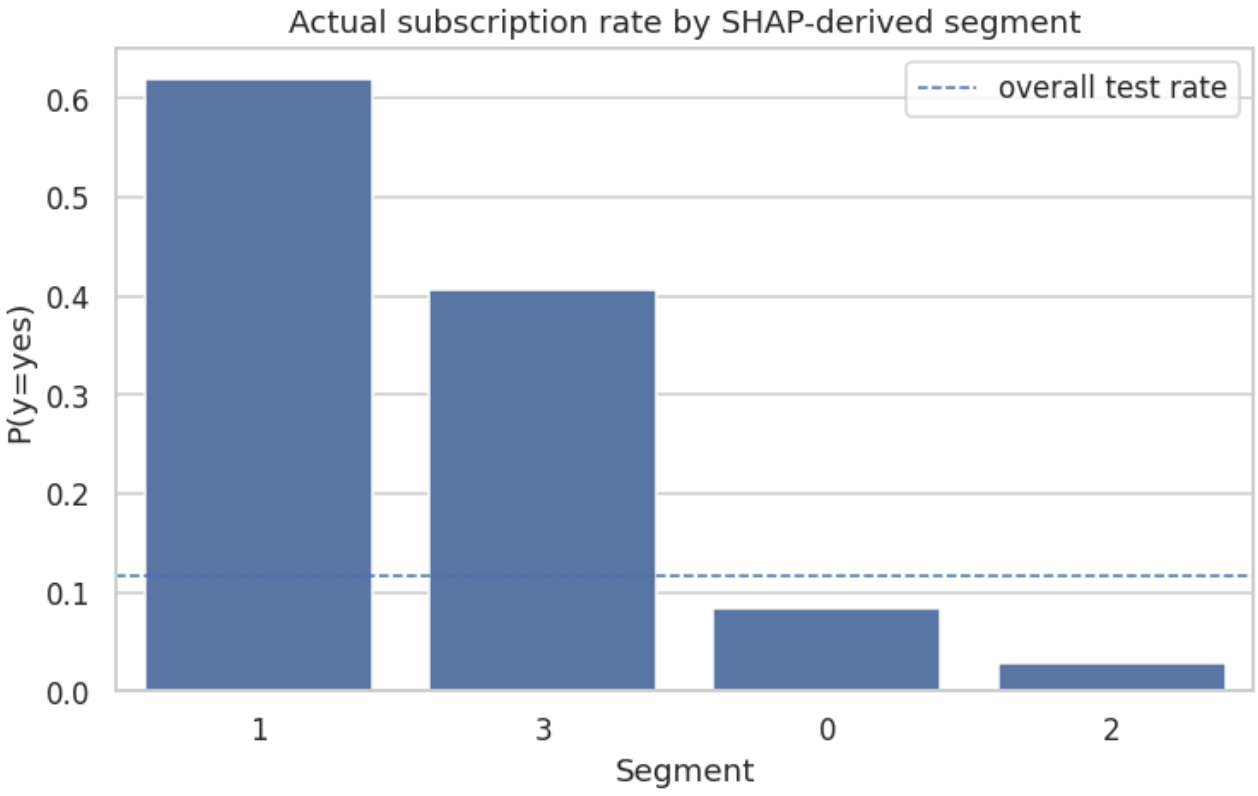

Figure 9: Subscription rates for four exploratory segments formed from SHAP explanation vectors

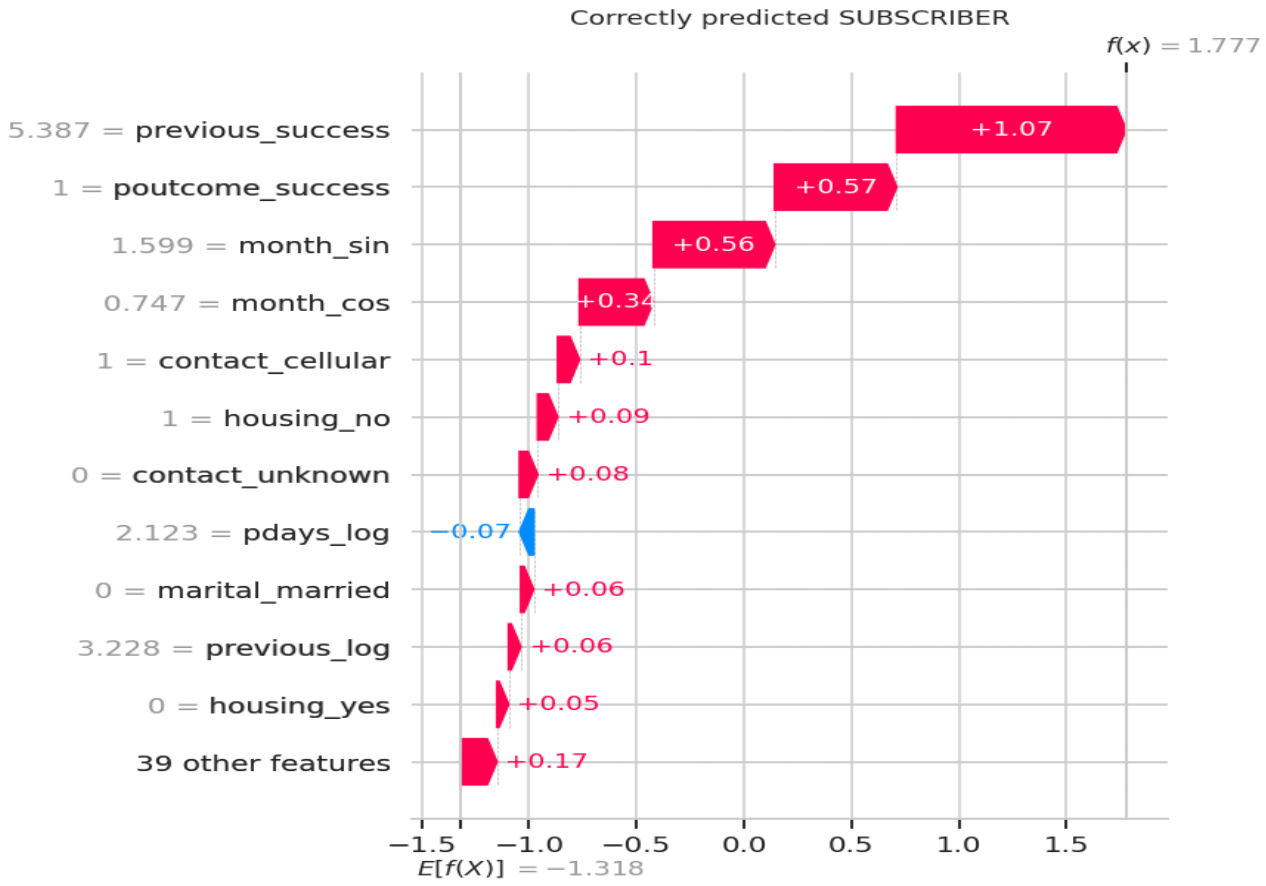

Figure 10: Supplied local SHAP waterfall for a correctly predicted subscriber

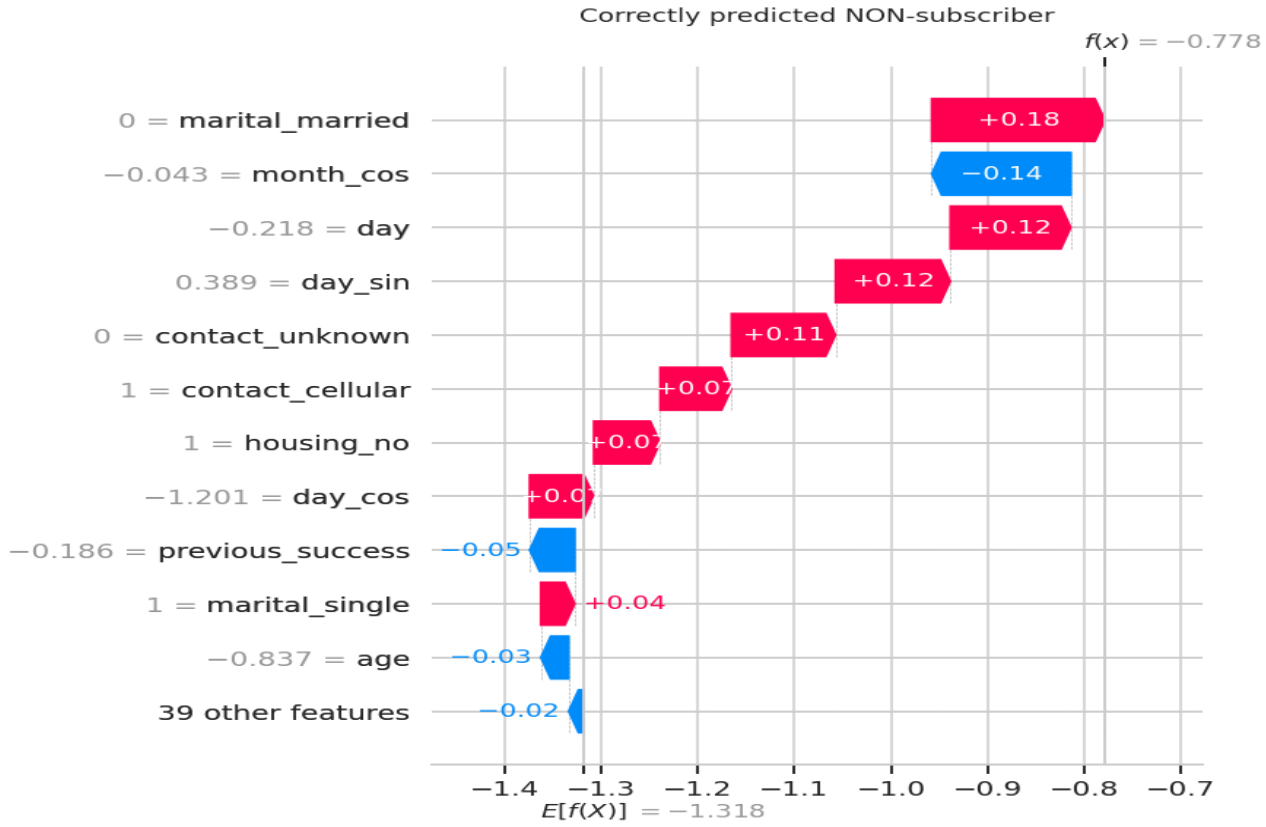

Figure 11: Supplied local SHAP waterfall for a correctly predicted non subscriber

Table 8: Subgroup fairness audit for the deployed LightGBM

Grouping variable	Selection rate ratio	TPR range	FPR range
Job category	0.087	0.273	0.100
Marital status	0.682	0.104	0.008
Age group	0.652	0.193	0.210

correlated pathways rather than the protected attributes directly. The result indicates that the classifier as currently trained should not be deployed for pre contact targeting without a fairness mitigation step, such as a constrained threshold per group, a post processing calibration adjustment, or a fairness aware training objective; the present study does not evaluate any such mitigation and reports the audit as a diagnostic finding.

8.7 Exploratory Extensions

The attention based FT Transformer extension obtains test ROC AUC 0.8014, average precision 0.4411, and F1 0.4496. A training time cost weighted LightGBM extension obtains ROC AUC 0.8089, average precision 0.4660, recall 0.7940, F1 0.3455, and expected profit 48,180 at the default threshold, compared with 16,595 for the calibrated LightGBM model at the same threshold; the cost weighted model trades precision for a substantially higher recall and a higher expected profit under the illustrative cost matrix, but it has not been calibrated and reuses the held out test set after the primary comparison, so its result is exploratory rather than confirmatory.

9 DISCUSSION

The largest observed performance change comes from information timing, not from the choice among three strong boosting libraries. Including duration increases realistic training set cross validated ROC AUC by approximately 0.14 for every tested family. Excluding duration lowers apparent discrimination but produces a model that can be used before contact. This result is consistent with the dataset documentation and with the earlier bank telemarketing literature, which distinguishes variables available for targeting from variables observed during call execution [17, 1].

The random test results show a narrow range among the tuned tree systems. The calibrated LightGBM is not the numerical winner on every metric. Stacking has the highest ROC AUC, voting has the highest average precision, and LightGBM has the highest F1 at the default threshold. Such differences support reporting a model selection rule rather than declaring a universal winner. The selection rule used here prioritizes training set cross validated average precision for the single calibrated model and reserves the test set for final reporting.

Calibration materially improves Brier score while leaving ranking nearly unchanged. This separation matters because a ranking model can identify high scoring records without producing probabilities suitable for expected profit arithmetic. The profit threshold is substantially below 0.5 because the assumed benefit of a successful subscription is larger than the cost of an unnecessary contact. Actual deployment requires institution specific costs, capacity constraints, and contact fatigue effects.

The temporal stress test is substantially harder than the random split. The later period has a higher subscription prevalence, and the available month field does not identify a calendar year. The resulting estimate should therefore be interpreted as evidence of temporal sensitivity rather than as a precise forecast of a future campaign. Exact timestamps and repeated campaign identifiers would be required for a stronger temporal design.

10 LIMITATIONS AND REPRODUCIBILITY

The evidence is limited to one public institution and one historical market context. Transfer to another institution, country, product, or period requires external validation. The chronological evaluation is inferred from month and day because the released file lacks an explicit year field in the analysis version. Hyperparameter optimization uses a 12,000 record stratified subsample for efficiency, although the selected models are refit on the full training partition. Full data nested optimization could produce different choices.

The cost parameters are illustrative and should not be interpreted as financial advice or a realized return estimate. The SHAP explanations are associative model explanations, not interventions. The exploratory FT Transformer and cost sensitive extensions reuse the held out test set after the primary candidate table, so their results require independent confirmation.

The paper is reproducible from the supplied notebook, the UCI dataset citation, the stated random split, the feature definitions, and the reported hyperparameter search settings. A future release should include an environment file, the exact random seed, raw result tables, confidence intervals for average precision and all operating metrics, and a fully nested evaluation for threshold and extension selection.

11 CONCLUSION

A credible bank subscription model must respect the time at which each variable becomes available. In the supplied UCI experiment, call duration creates a large and repeatable performance inflation and is therefore excluded from the deployable pipeline. After leakage control, tuned tree models achieve test ROC AUC close to 0.81, with average precision near 0.47 and performance that is sensitive to the operating threshold. Isotonic calibration improves probability quality, while a training set selected profit threshold increases expected test profit under an explicitly illustrative cost matrix. The temporal stress test reveals a marked reduction in discrimination, emphasizing the need for time aware validation. The principal practical conclusion is that validity, calibration, and decision policy should be treated as first class components of predictive modeling rather than as postscript additions to an accuracy comparison.

AUTHOR CONTRIBUTION STATEMENT

All authors contributed equally to the study conception and design. Material preparation, data collection, and analysis were performed by the authors. The first draft of the manuscript was written by the authors, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

This study did not involve human participants or animals. Ethical approval and consent to participate are therefore not applicable.

CONSENT FOR PUBLICATION

Not applicable.

DATA AVAILABILITY

The Bank Marketing dataset used in this study is publicly available from the UCI Machine Learning Repository at `https://archive.ics.uci.edu/dataset/222/bank+marketing` and is identified by DOI 10.24432/C5K306 [2].

ACKNOWLEDGMENTS

The authors acknowledge the use of ChatGPT-5 for assistance in refining the manuscript. The authors reviewed and edited all AI-generated output and take full responsibility for the final content.

FUNDING

This research received no external funding.

DISCLOSURE STATEMENT

The authors declare that they have no competing interests.

REFERENCES

[1] S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing," *Decision Support Systems*, vol. 62, pp. 22–31, 2014.

[2] S. Moro, P. Rita, and P. Cortez, "Bank marketing," 2014. Repository record donated in 2012; dataset citation issued in 2014.

[3] G. Marinakos and S. Daskalaki, "Imbalanced customer classification for bank direct marketing," *Journal of Marketing Analytics*, vol. 5, no. 1, pp. 14–30, 2017.

[4] S. C. K. Tékouabou, Ş. C. Gherghina, H. Toulni, P. Neves Mata, M. N. Mata, and J. M. Martins, "A machine learning framework towards bank telemarketing prediction," *Journal of Risk and Financial Management*, vol. 15, no. 6, p. 269, 2022.

[5] S. F. Crone, S. Lessmann, and R. Stahlbock, "The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing," *European Journal of Operational Research*, vol. 173, no. 3, pp. 781–800, 2006.

[6] E. Duman, Y. Ekinci, and A. Tanrıverdi, "Comparing alternative classifiers for database marketing: The case of imbalanced datasets," *Expert Systems with Applications*, vol. 39, no. 1, pp. 48–53, 2012.

[7] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[8] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[9] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," *Advances in neural information processing systems*, vol. 31, 2018.

[10] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," *Advances in neural information processing systems*, vol. 34, pp. 18932–18943, 2021.

[11] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PloS one*, vol. 10, no. 3, p. e0118432, 2015.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[13] A. Atwa, A. Y. Ismaeel, and A. A. Elngar, "Machine learning for chronic disease classification and comorbidity detection: Methodological gaps and future directions," *Journal of Smart Algorithms and Applications (JSAA)*, vol. 3, no. 2, pp. 87–104, 2026.

[14] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proceedings of the 22nd international conference on Machine learning*, pp. 625–632, 2005.

[15] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[16] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, no. Jan, pp. 1–30, 2006.

[17] S. Moro, P. Rita, and P. Cortez, "Bank marketing," 2014. Dataset documentation accessed August 26, 2026.

[18] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 20, no. 2, pp. 215–232, 1958.

[19] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

[20] C. Elkan, "The foundations of cost-sensitive learning," in *Proceedings of the International Joint Conference on Artificial Intelligence*, vol. 17, pp. 973–978, 2001.

[21] A. H. Murphy, "A new vector partition of the probability score," *Journal of Applied Meteorology and Climatology*, vol. 12, no. 4, pp. 595–600, 1973.

[22] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[23] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268, 2015.