applied sciences

MDPI

Article

A Composable Additive-Tree Hybrid with a Stabilised Decision Threshold for Explainable Churn Decision Support in Telecommunications Information Systems

Abdalrazag Beshti *, Oluwatayomi Rereloluwa Adegboye and Ahmad Bassam Alzubi

Department of Business Administration, Institute of Graduate Research and Studies, University of Mediterranean Karpasia, Mersin-10, Northern Cyprus, TR-10, Mersin 99010, Turkey
* Correspondence: 230635027@std.akun.edu.tr

Abstract

Customer churn prediction is a core predictive analytics function in telecommunications information systems, where model outputs feed retention campaigns and resource allocation decisions. Although supervised learning is widely used to rank subscribers by attrition risk, published comparisons rarely hold the tuning protocol, resampling design, and decision rule constant, so it remains unclear which part of the pipeline actually generates business value. This study addresses that gap on a widely used public benchmark of 7043 subscriber records described by 19 native attributes, with a churn prevalence of 26.54 per cent. Five learner families and a proposed Composable Additive-Tree Convex Hybrid (CATCH) were evaluated under a single nested protocol: repeated stratified five-fold cross-validation with three repeats, an 18-configuration random search scored by inner three-fold cross-validation, and a decision threshold estimated on the training partition alone. CATCH aggregates four learners whose outputs admit an exact additive decomposition on the logit scale, and couples them with a plateau-stabilised, metric-aware threshold estimator that translates risk scores into operational retention decisions. The hybrid obtained the highest mean area under the receiver operating characteristic curve (0.8499), F1 score (0.6380), and Matthews correlation coefficient (0.4926), exceeding the class-prior benchmark by a wide margin on every threshold metric, although its margins over the tuned single learners were small. Signed-rank tests indicated a ranking advantage over four of the five baselines, of which only that over the random forest persisted under a conservative corrected test, while the explainable boosting machine remained statistically indistinguishable; the hybrid is therefore presented as an exactly attributable aggregate. Decomposing the sources of improvement showed that component tuning and the decision layer dominate: replacing the conventional fixed threshold of 0.5 raised the Matthews correlation coefficient by 0.0217 for the hybrid and by as much as 0.0374 for the random forest, whereas aggregation contributed only 0.0005 in ranking quality. Composed Shapley attribution proved more stable than permutation importance and identified contract type, tenure, and internet service as the dominant drivers, giving retention managers interpretable evidence to act on. For organisations deploying churn models in enterprise systems, the results obtained on this benchmark indicate that well-structured tuning and threshold selection deliver more decision value than ensemble complexity.

Keywords: customer churn prediction; explainable machine learning; predictive analytics; data-driven decision making; decision threshold selection; metaheuristic optimization; decision support systems

Check for updates

Academic Editor: George Drosatos

Received: 10 September 2026
Revised: 15 October 2026
Accepted: 16 October 2026
Published: 24 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Telecommunications operators in mature markets grow almost entirely by retention. Once penetration approaches saturation, the marginal subscriber is won from a competitor rather than from an unserved population, and acquiring a new subscriber is costlier than retaining an existing one [1]. Attrition therefore functions as a direct tax on recurring revenue, and the ability to identify which subscribers are at risk while an intervention is still possible has become a routine requirement of commercial analytics rather than a research curiosity. Churn is studied in banking as well as in telecommunications, and churn modelling has become a common application of supervised learning on tabular data [2,3].

The methodological history of the problem is short and largely additive. Early treatments applied logistic regression and decision trees to operator billing records. The subsequent decade replaced them with ensembles of trees, then with deep architectures, and most recently with post hoc explanation layers intended to convert a risk score into a retention action [4–6]. Each wave has reported improvements over the last. What has not kept pace is the comparability of those reports. Studies differ in the dataset, in the treatment of class imbalance, in whether hyperparameters are selected inside or outside the evaluation loop, and in the threshold at which a probability is converted into a retention decision. Two recent reviews, one covering 212 published studies and the other 240 studies in a bibliometric analysis, likewise identify the profit-oriented, business-aligned evaluation of churn models as an under-researched part of the modelling process [7,8].

This matters because the practical question an operator faces is not which family of learners is fashionable but where an additional unit of modelling effort should be spent. Three candidate destinations present themselves: enlarging the hypothesis class, tuning the chosen class more thoroughly, or improving the rule that turns a calibrated score into a contact list. The literature has concentrated overwhelmingly on the first. The second is usually reported without an untuned reference point, so the size of the tuning contribution cannot be recovered. The third is very often left at the library default, a value that has no particular claim on a problem whose positive class occupies roughly a quarter of the sample, and on which threshold-dependent summaries such as accuracy and the F1 score can give an over-optimistic picture [9,10]. Two gaps follow. First, to the best of our knowledge, no published comparison of the principal tabular learner families on this task holds the resampling design, the search budget, the search objective, and the decision rule simultaneously fixed, which is the only arrangement under which a difference between learners can be attributed to the learners. Second, the relative size of the ranking contribution and the decision contribution has not been quantified on a common footing, even though the two are routinely conflated in a single headline metric.

The present study addresses both. Its objective is to determine, under one nested protocol applied identically to every candidate, whether a proposed aggregation of complementary learners improves on the strongest individual learner, and to decompose whatever improvement is obtained into the part originating in the ranking layer and the part originating in the decision layer. The proposed framework, referred to as the Composable Additive-Tree Convex Hybrid (CATCH), combines four learners in the space of log-odds. Its members are chosen so that the aggregate inherits an exact additive attribution from its components, and its decision threshold is estimated from a stabilised neighbourhood of the optimum on the training partition rather than from a single maximiser. Evaluation proceeds by repeated stratified five-fold cross-validation on 7043 subscriber records described by 19 native attributes, and reports the area under the receiver operating characteristic curve, the F1 score, accuracy, balanced accuracy, and the Matthews correlation coefficient against a class-prior benchmark and five tuned baselines.

The contributions of the manuscript are as follows. The first concerns evaluation: a single nested protocol, with an untuned reference point and a class-prior benchmark, under which the contributions of the hypothesis class, of tuning and of the operating point can be measured on a common footing. The second is the decision layer, a plateau-stabilised, metric-aware threshold estimator that is independent of the learner and accounts for most of the measured improvement. The third is the hybrid, whose novelty lies neither in its members nor in a gain in predictive accuracy, but in restricting a convex mixture on the logit scale to members with exact additive decompositions, so that the aggregate carries an exact attribution. The fourth is a fold-level comparison of the stability of two attribution methods. Tuning and threshold selection apply equally to every baseline and are reported as findings about the methodology. The methodological contribution is therefore not a further learner but two components: a decision layer that can be attached to any learner and that, on average across the six models, improves on both the fixed threshold and the single maximiser of the criterion curve on the F1 score, balanced accuracy and the Matthews correlation coefficient (Section 5.4), and an aggregation whose attribution is exact and more stable across folds than a model-agnostic alternative (Section 5.6).

The remainder of the manuscript is organised as follows. Section 2 reviews recent empirical work on churn prediction and on the evaluation of imbalanced classifiers. Section 3 presents the proposed framework and each of its components. Section 4 specifies the experimental protocol. Section 5 reports and discusses the outcomes, including the attribution analysis, a comparison with previous studies, and the operational consequences. Section 6 concludes.

2. Related Work

Ahmad et al. constructed a churn model for a Syrian mobile operator on a distributed processing platform, enriching conventional billing descriptors with social-network features derived from call graphs; the authors reported that adding the network descriptors raised the area under the curve from 84 to 93.3 per cent [4]. Óskarsdóttir et al. [11] benchmarked relational learners on call-network data and likewise found that the best models are conventional classifiers enriched with network features. Lalwani et al. [3] assembled a pipeline combining gravitational-search feature selection with several classifiers and boosting ensembles on a telecommunications dataset, and reported that adaptive boosting and extreme gradient boosting gave the highest accuracy and shared the highest area under the curve. Guliyev and Tatoğlu [2] transposed the problem to retail banking, pairing extreme gradient boosting with Shapley additive explanations to expose the contribution of loan and customer descriptors to attrition risk. Khattak et al. [5] developed a composite deep architecture for churn, placing a bidirectional recurrent layer ahead of convolutional feature extraction, and demonstrated gains over single-branch neural baselines on a public dataset. Chang et al. [12] compared several classifiers on a public telecommunications dataset of 7043 customers and 38 attributes and coupled the comparison with explainable artificial intelligence techniques, with the random forest performing best.

Poudel et al. [13] compared several tabular learners for telecommunications churn, found gradient boosting the most accurate under a Wilcoxon signed-rank comparison, and concentrated on explaining rather than merely maximising the predictions through Shapley additive explanations. Asif et al. [14] advanced a boosting ensemble combined with explainable artificial intelligence for telecommunications attrition, integrating several gradient-boosting variants through a soft-voting mechanism and reporting the contribution of individual descriptors alongside the aggregate performance. Christopher and Anand [15] compared eleven classifiers, among them random forests, extreme gradient boosting, logistic regression, and neural networks, found logistic regression the strongest,

and combined the predictive stage with a clustering-based segmentation of the customer base, an arrangement that separates who is at risk from what should be offered to them. Hao [16] introduced a hybrid sequence model in which a bidirectional gated recurrent unit with attention supplies representations to an extreme gradient-boosting classifier, and reported that the hybrid surpassed conventional machine-learning and deep-learning classifiers on an open-source dataset. El Attar and El-Hajj [6] implemented a seven-model comparison on the same 7043-record public benchmark used in this study, with engineered features and synthetic minority oversampling; gradient-boosting variants achieved balanced performance near 0.84 on accuracy, precision, recall, and the F1 score, extreme gradient boosting attained the strongest discrimination with an area under the receiver operating characteristic curve of 0.932, and a soft-voting ensemble of the leading models did not exceed it, with an F1 score of 0.84 and an area under the curve of 0.918.

Larger benchmarking exercises place these single-dataset comparisons in context. Verbeke et al. [17] evaluated a broad set of classifiers on eleven datasets from telecommunications operators and found that a large group of them performs comparably and that oversampling does not in general improve performance significantly. Coussement et al. [18] showed that a logistic regression with careful data preparation is competitive with more advanced learners, and Geiler et al. [19] benchmarked eleven learning methods and seven sampling approaches on sixteen churn-like datasets. Outside the churn literature, Hand [20] argued that the improvements attributed to sophisticated classifiers are frequently smaller than the uncertainties of deployment, tabular benchmarks continue to find tuned tree ensembles at least as strong as deep architectures [21,22], and Probst et al. [23] quantified, on 38 datasets and six algorithms, how much of the performance of an algorithm depends on the tuning of its hyperparameters.

Hybrid and interpretable constructions have a long history in churn modelling. De Bock and Van den Poel [24] proposed ensembles of generalised additive models to reconcile performance with interpretability, De Caigny et al. [25] combined decision trees with logistic regression, and Rudin [26] argued that models interpretable by construction are preferable to post hoc explanations of opaque ones for high-stakes decisions. On the aggregation side, stacked generalisation [27] and its variant with non-negative combining weights [28] are the standard devices, while the forecast-combination literature has shown that estimated optimal weights often fail to improve on equal weights [29]. The hybrid proposed here differs from these constructions in one respect: its pool is restricted to members whose log-odds decompose exactly into additive attribute contributions, so that the attribution of the aggregate is exact.

Turning to evaluation rather than modelling, Chicco and Jurman [10] argued that accuracy and the F1 score present over-optimistic pictures on imbalanced confusion matrices, and advocated the Matthews correlation coefficient as a summary that responds to all four cells; they later extended the argument in favour of the coefficient over the area under the receiver operating characteristic curve for assessing binary classifiers [30]. Zou et al. [9] demonstrated that the conventional cut-off of 0.5 is usually unsuitable under imbalance and proposed selecting the threshold explicitly, obtaining substantial gains in F-score. Holzmann and Klar [31] showed that several established summaries, the Matthews correlation coefficient among them, degrade towards classifiers that ignore the minority class as prevalence falls, and proposed robust modifications. Lipton et al. [32] showed that, for calibrated probabilities, the threshold that maximises the F1 score equals half of the optimal F1 score and therefore lies at or below 0.5, and Elkan [33] that the cost-optimal threshold follows directly from the entries of the misclassification cost matrix. Van den Goorbergh et al. [34] found that imbalance corrections such as oversampling distort calibration without improving discrimination, and that shifting the threshold achieves the same

classification benefit. Cawley and Talbot [35] showed that selecting hyperparameters on data later used for evaluation biases the reported performance, which motivates the nested protocol adopted here. Taken together, this strand argues that the choice of operating point deserves the same methodological care usually reserved for the learner, an argument the present study makes quantitatively.

A further strand evaluates churn models by the value of the campaign they support. Verbeke et al. [17] and Verbraken et al. [36] introduced maximum-profit and expected-maximum-profit criteria, respectively, and Ascarza [37] and Devriendt et al. [38] showed that targeting by predicted risk can be inferior to targeting by the predicted response to the intervention. These approaches presuppose cost, benefit, or treatment information that public benchmarks do not carry, and they act on the same step from score to decision that the decision layer of Section 3.4 addresses with statistical criteria.

The gap that remains concerns design and not method. The studies reviewed above differ in the dataset, in the treatment of class imbalance, in whether hyperparameters are tuned, and in the decision threshold, and two recent reviews identify business-aligned evaluation as a remaining weakness of the field [7,8]. The multi-dataset benchmarks [17,19] do not decompose performance into the contributions of tuning and of the operating point, while the threshold-selection literature [9,32,34] treats the operating point separately from the choice of learner. To the best of our knowledge, therefore, no study of churn prediction holds the resampling design, the search budget, the search objective, and the decision rule fixed across learner families, so that the contributions of the hypothesis class, of tuning, and of the operating point can be separated and compared. That is the gap addressed here.

3. Proposed Framework

This section sets out the framework in the order in which a prediction is produced: the overview, data, and preparation (Section 3.1), the component learners (Section 3.2), the aggregation and the decision layer (Sections 3.3 and 3.4), which are the two elements specific to the framework, and the evaluation criteria, search strategies, and attribution methods (Sections 3.5–3.7).

3.1. Overview, Data and Preparation

The framework is organised as the six-stage procedure shown in Figure 1. Stage 0 prepares the attributes; Stage 1 establishes the outer resampling design; Stage 2 tunes each component learner using only the training partition of the current outer fold; Stage 3 aggregates the components in the space of log-odds; Stage 4 converts the aggregated score into a decision; and Stage 5 evaluates the result and attributes it to the input attributes. No quantity used in Stages 2 to 4 is estimated on material that later serves as a test partition.

The data comprise 7043 subscriber records, each described by the 19 native attributes listed in Table 1 and labelled by a binary churn indicator. The positive class accounts for 26.54 per cent of the records. The data are publicly available on GitHub [39]. Three attributes are numeric and sixteen are categorical; of the categorical attributes, six are binary and ten have three or four levels. The total charges attribute is stored as a text field in the dataset and is blank for the eleven records whose tenure is zero; those entries were recoded as zero, the value implied by a subscriber who has not yet completed a billing cycle. No other imputation, resampling, or synthetic augmentation was applied, so the class distribution presented to every learner is the one observed in the data.

Numeric attributes were standardised, and, where a component learner requires it, expanded on a cubic B-spline basis so that a linear model could represent a curved association without abandoning additivity. Categorical attributes were encoded by indicator variables with an unseen-level fallback. Both encodings were fitted on the training partition only.

Figure 2 summarises the association between the principal attributes and the outcome: attrition concentrates in month-to-month agreements and fibre-optic connections, declines steeply with tenure, and separates only partially on monthly charges.

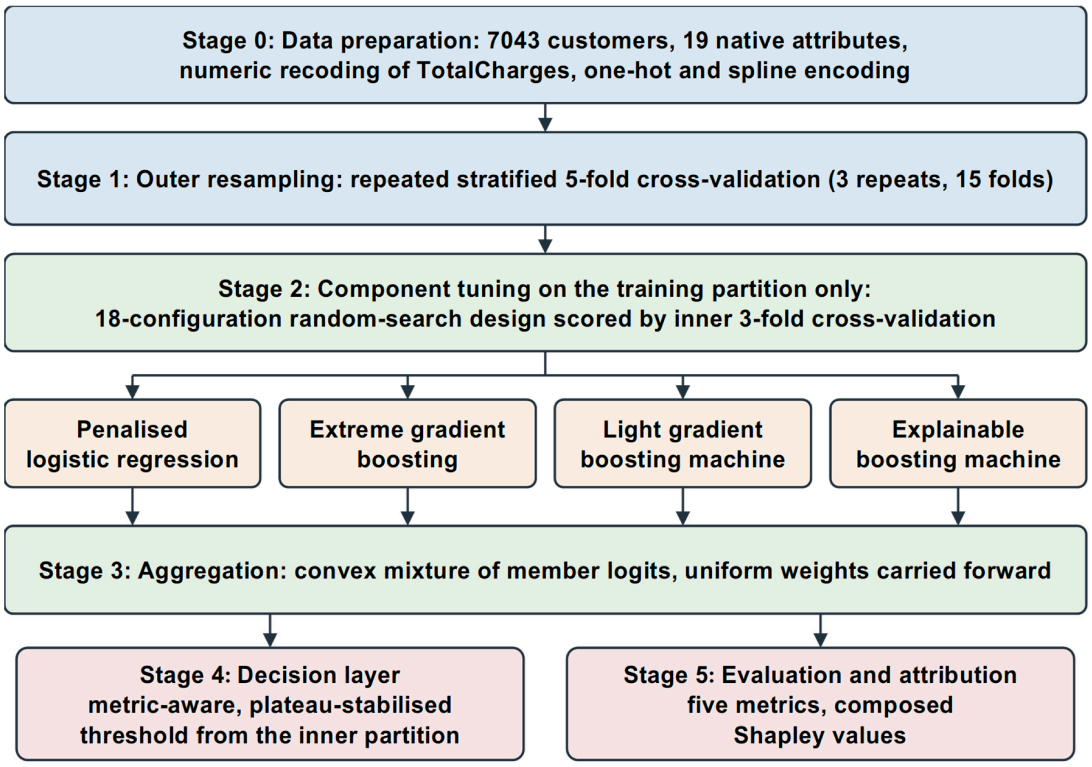
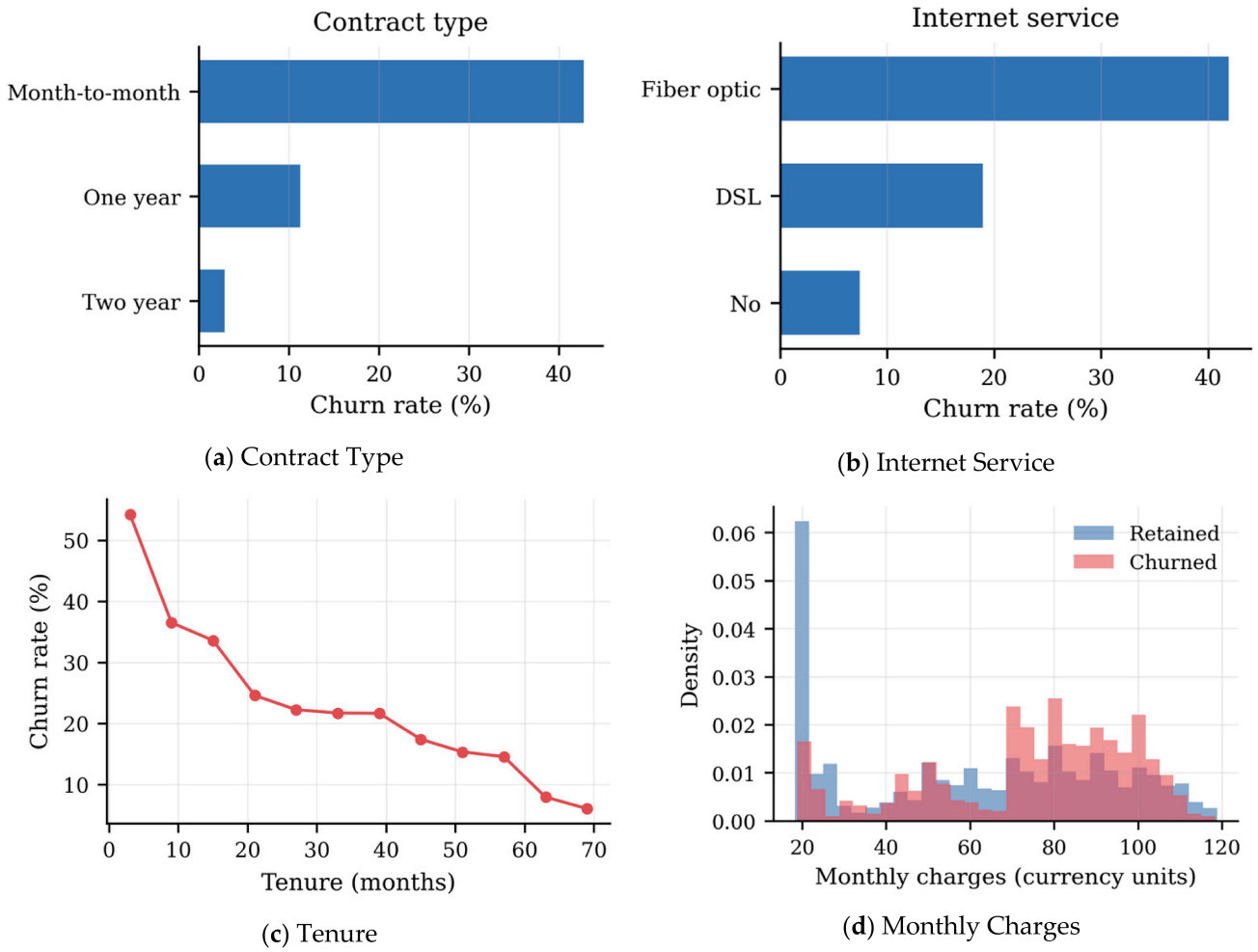


Figure 1. The proposed Composable Additive-Tree Convex Hybrid (CATCH) framework, organised as a six-stage pipeline from data preparation (Stage 0) to evaluation and attribution (Stage 5).

Table 1. The 19 native subscriber attributes.

Panel A: numeric attributes						
Attribute	**Type**	**Mean**	**Std. dev.**	**Minimum**	**Maximum**	**Correlation**
tenure	Numeric	32.37	24.56	0.00	72.00	−0.352
MonthlyCharges	Numeric	64.76	30.09	18.25	118.75	+0.193
TotalCharges	Numeric	2279.73	2266.79	0.00	8684.80	−0.198
Panel B: categorical attributes						
Attribute	**Levels**	**Modal share (%)**		**Min. churn (%)**	**Max. churn (%)**	**Cramér's *V***
gender	2	50.5		26.2	26.9	0.008
SeniorCitizen	2	83.8		23.6	41.7	0.150
Partner	2	51.7		19.7	33.0	0.150
Dependents	2	70.0		15.5	31.3	0.164
PhoneService	2	90.3		24.9	26.7	0.011
PaperlessBilling	2	59.2		16.3	33.6	0.191
MultipleLines	3	48.1		24.9	28.6	0.040
InternetService	3	44.0		7.4	41.9	0.322
OnlineSecurity	3	49.7		7.4	41.8	0.347
OnlineBackup	3	43.8		7.4	39.9	0.292
DeviceProtection	3	43.9		7.4	39.1	0.282
TechSupport	3	49.3		7.4	41.6	0.343
StreamingTV	3	39.9		7.4	33.5	0.231
StreamingMovies	3	39.5		7.4	33.7	0.231
Contract	3	55.0		2.8	42.7	0.410
PaymentMethod	4	33.6		15.2	45.3	0.303



(**a**) Contract Type

(**b**) Internet Service

(**c**) Tenure

(**d**) Monthly Charges

Figure 2. Attribute-outcome associations. (**a**) Churn rate by contract type; (**b**) churn rate by internet service; (**c**) churn rate against tenure in six-month bands (the last band, 66 to 72 months, includes the maximum tenure); (**d**) distribution of monthly charges for retained and churned subscribers.

Let $\mathcal{D} = \{(\mathrm{x}_i, y_i)\}_{i=1}^{n}$ denote the sample, with $x_i \in \mathcal{X}$ the vector of encoded attributes of subscriber i,$y_i \in \{0, 1\}$ the churn indicator, and $n = 7043$. Every component learner produces an estimate of the conditional probability

$$p(\mathrm{x}) = \Pr(y = 1 \mid \mathrm{x}), \tag{1}$$

where p:$\mathcal{X} = (0, 1)$. Equation (1) defines the common output on which all components, and hence the aggregate, are compared. It is convenient to work with the log-odds transform

$$z(\mathrm{x}) = \log \frac{p(\mathrm{x})}{1 - p(\mathrm{x})}, \tag{2}$$

in which z:$\mathcal{X} = \mathbb{R}$ is unbounded and additive contributions compose linearly. Equation (2) is the scale on which the aggregation of Section 3.3 and the attribution of Section 3.7 are both defined.

3.2. Component Learners

The five families below span the inductive biases commonly applied to the dataset: a strictly additive parametric model, a bagged high-variance ensemble, two sequentially boosted tree ensembles with differing splitting policies, and a boosted generalised additive

model. They are established methods, so only the property on which the framework relies is stated for each.

3.2.1. Penalised Logistic Regression (PLR)

The first component, PLR, models the log-odds as a linear function of the encoded attributes. Writing $\beta \in \mathbb{R}^d$ for the coefficient vector over the d encoded columns and β_0 for the intercept, the fitted coefficients minimise the penalised negative log-likelihood

$$\mathcal{L}(\beta, \beta_0) = -\sum_{i=1}^{m} [y_i \log p(\mathrm{x}_i) + (1 - y_i)\log(1 - p(\mathrm{x}_i))] + \frac{1}{2C}\|\beta\|_2^2, \tag{3}$$

where m is the number of training records, $\| \cdot \|_2$ is the Euclidean norm, and $C > 0$ is the inverse regularisation strength. Equation (3) yields a model whose contribution to the log-odds is exactly linear in each encoded column, which is what later permits an exact attribution without approximation. The family is included because a strictly additive model is a demanding reference point on subscriber data, and because its errors are structurally different from those of a tree ensemble.

3.2.2. Random Forest (RF)

The second component, RF, averages the outputs of B classification trees, each grown on a bootstrap resample and each restricted at every split to a random subset of the encoded columns [40], and, as implemented in scikit-learn, takes the mean of the class-one leaf frequencies of the trees as the ensemble estimate. The averaging reduces the variance of a single deep tree while leaving its low bias largely intact. The family is included as the canonical bagged ensemble and as the strongest available contrast to sequential boosting.

3.2.3. Extreme Gradient Boosting

The third component builds an additive expansion of regression trees, fitting each new tree to a second-order approximation of the loss and penalising tree complexity explicitly [41]. After T rounds the log-odds are

$$z_{\mathrm{XGB}}(\mathbf{x}) = \sum_{t=1}^{T} \eta f_t(\mathrm{x}), f_t \in \mathcal{F}, \tag{4}$$

where $\eta \in (0,1]$ is the learning rate, f_t is the regression tree fitted at round t, and $\mathcal{F}$ is the space of such trees. The regularised second-order objective minimised at each round is given in the original description of the method and is not repeated here. Equation (4) makes the log-odds an additive function of trees, which is what later allows an exact tree-based attribution.

3.2.4. Light Gradient Boosting Machine

The fourth component follows the same additive scheme but grows each tree leaf-wise, expanding the leaf with the largest loss reduction rather than completing each depth level, and bins continuous attributes into histograms before searching for split points [42]. The leaf-wise policy produces deeper, more asymmetric trees than the level-wise policy of the previous component, and the family is therefore included to represent a second and materially different boosting strategy; the split-gain criterion is that of the original method.

3.2.5. Explainable Boosting Machine (EBM)

The fifth component, EBM, restricts the expansion to shape functions of individual attributes, together with an optional set of pairwise terms, and fits them by cyclic gradient boosting with very small steps [43,44]. Its log-odds are

$$z_{\mathrm{EBM}}(\mathrm{x}) = \beta_0 + \sum_{j=1}^{q} s_j(x_j) + \sum_{(j,k)\in\mathcal{P}} s_{jk}(x_j, x_k), \tag{5}$$

where q is the number of native attributes, s_j is the univariate shape function of attribute j,$\mathcal{P}$ is the selected set of attribute pairs, and s_{jk} is the corresponding bivariate shape function. Equation (5) is additive by construction, so each term is its own exact attribution. The family is included both as a strong tabular learner and because it supplies the additive structure on which the interpretability of the aggregate rests.

3.3. Aggregation

The four components whose log-odds decompose exactly into additive attribute contributions, namely penalised logistic regression, extreme gradient boosting, the light gradient boosting machine, and the explainable boosting machine, form the aggregation pool. Denoting their log-odds by $z_k(\mathrm{x})$ for $k = 1,\ldots,K$ with $K = 4$, the hybrid predicts

$$z_{\mathrm{CATCH}}(\mathrm{x}) = \sum_{k=1}^{K} w_k z_k(\mathrm{x}),\ p_{\mathrm{CATCH}}(\mathrm{x}) = \left(1 + e^{-z_{\mathrm{CATCH}}(\mathrm{x})}\right)^{-1}, \tag{6}$$

subject to the simplex constraint

$$\sum_{k=1}^{K} w_k = 1, w_k \geq 0 \forall k \tag{7}$$

where w_k is the weight assigned to component k. Equation (6) combines the members on the scale defined by Equation (2), and Equation (7) keeps the mixture convex, which prevents the extrapolation that an unconstrained meta-learner can produce when its inputs are strongly correlated. Because the members are highly correlated, estimating w freely is prone to variance. The framework therefore shrinks the estimated weight vector towards the uniform point of the simplex,

$$\mathrm{w}_\gamma = (1-\gamma)\hat{\mathrm{w}} + \gamma\frac{1}{K}1, \gamma \in [0,1], \tag{8}$$

with $\hat{\mathrm{w}}$ the weight vector obtained by search, 1 the vector of ones and γ the shrinkage coefficient selected on the inner partition by a one-standard-error rule. Equation (8) spans the continuum between a fully estimated mixture and a plain average, so the data determine how much estimation the aggregation can support. Weights are selected against a coupled criterion that combines the ranking quality of the mixture with the quality of the decision it supports,

$$J(\mathrm{w}) = \lambda \mathrm{AUC}(\mathrm{w}) + (1-\lambda)\max_{\tau} \mathrm{MCC}(\mathrm{w}, \tau), \tag{9}$$

in which AUC is the area under the receiver operating characteristic curve, MCC is the Matthews correlation coefficient, τ is the decision threshold, and $\lambda \in [0, 1]$ balances the two terms and was set to 0.5. Equation (9) is evaluated exclusively on out-of-fold predictions from the inner partition. The aggregation and the decision layer of Section 3.4 are the two elements specific to the proposed framework; component tuning, by contrast, is applied identically to every baseline.

3.4. Decision Layer

A probability becomes a retention decision through a threshold τ, and the value of every threshold-dependent metric is a property of the pair (p,τ) rather than of p alone. The framework estimates τ separately for each target metric on the inner out-of-fold predictions. Here, $M(\tau)$ denotes the criterion curve of the target metric evaluated over a grid $\mathcal{T}$ of 171 candidate thresholds spanning [0.05, 0.90]. Rather than adopting the single maximiser, which is a noisy statistic on a step function, the estimator averages the near-optimal plateau,

$$\hat{\tau} = \frac{1}{|\mathcal{T}_\epsilon|} \sum_{\tau \in \mathcal{T}_\epsilon} \tau, \mathcal{T}_\epsilon = \left\{ \tau \in \mathcal{T} : M(\tau) \geq (1-\epsilon) \max_{\tau' \in \mathcal{T}} M(\tau') \right\}, \tag{10}$$

where $\mathcal{T}_\epsilon$ is the set of thresholds within a relative tolerance ϵ of the maximum, $|\cdot|$ denotes cardinality, and ϵ is fixed at 0.01. Equation (10) returns the centre of the region in which the criterion is indistinguishable from its best value, which transfers more reliably to unseen records than the argument of the maximum. Because the estimator takes the criterion curve as its only input, a cost-weighted or profit-based criterion can be substituted for the target metric without altering the rest of the framework.

3.5. Evaluation Criteria

Five criteria are reported. The first is threshold-free; the remaining four are evaluated at the threshold supplied by Equation (10). With TP, FP, FN, and TN denoting the counts of true positives, false positives, false negatives, and true negatives at a given threshold, and $\mathcal{P}^+$ and $\mathcal{P}^-$ the sets of positive and negative records,

$$\text{AUC} = \frac{1}{|\mathcal{P}^+||\mathcal{P}^-|} \sum_{i \in \mathcal{P}^+} \sum_{j \in \mathcal{P}^-} \left[\mathbb{I}(p_i > p_j) + \frac{1}{2} \mathbb{I}(p_i = p_j) \right], \tag{11}$$

where $\mathbb{I}(\cdot)$ is the indicator function and p_i is the predicted probability for record i. Equation (11) estimates the probability that a churning subscriber is ranked above a retained one. The F1 score is the harmonic mean of precision and recall on the positive class, accuracy is the overall agreement rate, and balanced accuracy is the mean of the two class-conditional recalls; all three are used in their standard forms. The Matthews correlation coefficient, which also enters the coupled criterion of Equation (9), is

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}. \tag{12}$$

Equation (12) is a correlation between predicted and observed labels that responds to all four cells of the confusion matrix [10].

3.6. Search Strategies

Two search strategies were applied to the aggregation problem under an identical evaluation budget, so that the contribution of the search itself could be assessed.

3.6.1. Random Search

Configurations were drawn independently from the prior over the search space and scored by Equation (9), with the best-scoring configuration retained. Random search is used both for the component hyperparameters and as the reference strategy for the aggregation weights; its efficiency on low-effective-dimension spaces is well established [45].

3.6.2. Arithmetic Optimization Algorithm

The second strategy is the Arithmetic Optimization Algorithm, a population method in which the four elementary arithmetic operators govern the search: division and multiplication generate widely dispersed exploration moves, while subtraction and addition generate dense exploitation moves [46]. Two coefficients schedule the transition between the two regimes. The math optimiser accelerated function is

$$\mathrm{MOA}(t) = \mathrm{MOA}_{\min} + t\frac{\mathrm{MOA}_{\max} - \mathrm{MOA}_{\min}}{T}, \tag{13}$$

and the math optimiser probability is

$$\mathrm{MOP}(t) = 1 - \frac{t^{1/\alpha}}{T^{1/\alpha}}, \tag{14}$$

where t is the current iteration, T is the iteration budget, $\mathrm{MOA}_{\min}$ = 0.2 and $\mathrm{MOA}_{\max}$ = 1.0 bound the accelerated function, and α = 5 controls the decay of the exploitation accuracy. Equations (13) and (14) schedule the exploration (division and multiplication) and exploitation (subtraction and addition) moves of the original algorithm, which was implemented with the update rules and exploration condition of the published description and with the parameter values of the authors' reference implementation, namely an upper bound of 1.0 for the accelerated function and a control parameter of 0.499, against 0.9 and 0.5 in the published description. Before it was applied to the aggregation problem, the optimiser was verified on five standard benchmark functions in 30 dimensions, and the outcome is reported in Section 5.5.

3.7. Attribution Methods

Two candidate attribution methods were compared on the model that the results identify as best, and the more stable of the two was carried into the interpretation.

Composed Shapley attribution. For a prediction function v acting on a set $N = \{1,\ldots,q\}$ of attributes, the Shapley value assigned to attribute j is

$$\phi_j(v) = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(q-|S|-1)!}{q!}[v(S \cup \{j\}) - v(S)], \tag{15}$$

where S is a coalition of attributes excluding j, $|S|$ its cardinality, and $v(S)$ the model output when only the attributes in S are present [47,48]. Equation (15) is linear in v, and because the hybrid is a fixed linear combination of member log-odds by Equation (6), the attribution of the aggregate follows exactly from those of its members,

$$\phi_j(z_{\mathrm{CATCH}}) = \sum_{k=1}^{K} w_k \phi_j(z_k). \tag{16}$$

Equation (16) is the reason the aggregation pool was restricted to members with exact additive decompositions: the coefficients of the penalised logistic model give ϕ_j in closed form, tree ensembles admit the exact polynomial-time tree algorithm on the log-odds scale, and the shape functions of the explainable boosting machine are themselves the attributions. Contributions of encoded columns are summed back to their parent attribute, and a bivariate term is divided equally between the two attributes it involves. Global importance is the mean absolute value of ϕ_j over the held-out records.

Permutation importance. The second method is model-agnostic and measures the deterioration of ranking quality when the association between one attribute and the outcome is destroyed,

$$\mathrm{PI}_j = \mathrm{AUC}(p, y) - \frac{1}{R}\sum_{r=1}^{R} \mathrm{AUC}\left(p^{(j,r)}, y\right), \tag{17}$$

where $p^{(j,r)}$ is the prediction obtained after the r-th random permutation of attribute j across records and $R = 10$ is the number of repetitions; the measure is an AUC-based, repeated-permutation form of the model reliance of Fisher et al. [49]. Equation (17) attributes to each attribute the ranking quality that depends on it, and, unlike Equation (16), it requires no assumption about the internal form of the model.

4. Experimental Design

This section specifies the conditions of the comparison: the resampling protocol, the reference points and the search settings (Section 4.1), followed by the statistical assessment and the implementation (Section 4.2).

4.1. Resampling Protocol, Reference Points and Search Settings

The outer design is repeated stratified five-fold cross-validation with three repeats, giving 15 outer folds, each with 5634 training and 1409 test records and the class proportion of the full sample preserved in both. Within the training partition of each outer fold, a stratified three-fold inner split supplies the out-of-fold predictions on which hyperparameters, aggregation weights, the shrinkage coefficient, and the decision thresholds are all selected. The five component families and the hybrid therefore receive an identical protocol, an identical search budget, and an identical decision rule, so that any difference between them is attributable to the hypothesis class.

Two reference points frame the comparison. The class-prior benchmark assigns every subscriber the training prevalence, which yields an area under the receiver operating characteristic curve of exactly 0.5 and, at any threshold above the prevalence, the majority-class decision. The second reference point is each family's library-default configuration, evaluated under the same protocol, which isolates the contribution of tuning.

Table 2 records the settings of every search performed. Component hyperparameters were selected from a fixed random-search design of 18 configurations per family, identical across folds, that always includes the library default so that the tuner is never denied the default. Aggregation weights were searched by random search, by the Arithmetic Optimization Algorithm and by a Nelder–Mead simplex, each with the budget stated. The decision threshold was obtained by exhaustive evaluation of the 171-point grid followed by Equation (10). The objective J is defined in Equation (9); k denotes the outer fold index.

Table 2. Settings of every search performed in the study.

Stage	Strategy	Dimensions	Budget	Validation Scheme	Objective	Stopping Criterion	Seed
Component hyperparameters	Random-search design	5.2 (mean per family)	18 configurations	Inner 3-fold cross-validation	$J = \frac{1}{2}$ AUC + $\frac{1}{2}$ MCC	Exhausted design	2026
Aggregation weights	Random search	4	390 evaluations	Inner out-of-fold predictions	J	Exhausted budget	20260908 + k
Aggregation weights	Arithmetic Optimization Algorithm	4	390 evaluations (15 agents, 25 iterations)	Inner out-of-fold predictions	J	Iteration limit	20260908 + k

Table 2. *Cont.*

Stage	Strategy	Dimensions	Budget	Validation Scheme	Objective	Stopping Criterion	Seed
Aggregation weights	Nelder-Mead simplex	4	600 iterations	Inner out-of-fold predictions	J	Tolerance 10^{-8}	deterministic
Shrinkage coefficient	Grid with one-standard-error rule	1	6 values	Repeated 5-fold on inner predictions	J	Exhausted grid	20260908 + k
Decision threshold	Plateau-stabilised grid	1	171 values	Inner out-of-fold predictions	Target metric	Exhausted grid	deterministic
Family		**Search Space**					
Penalised logistic regression		C ∈ [0.01 ǀ 20] (log); numeric_transform ∈ {scale, spline }; n_knots ∈ {4,5,6,8}					
Random forest		n_estimators ∈{100,150,250}; max_depth ∈{6,8,12,16, None}; min_samples_leaf ∈{2,5,10,20,40}; max_features ∈{sqrt ,0.3,0.5}					
Extreme gradient boosting		n_estimators ∈{200,400,600}; learning_rate ∈[0.01,0.2](log); max_depth ∈{2,3,4,5, 6}; min_child_weight ∈{1,3,5,10,20}; subsample ∈{0.7,0.85,1.0}; colsample_bytree ∈{0.5,0.7,0.9}; reg_lambda ∈[0.5,50] (log)					
Light gradient boosting machine		n_estimators ∈{200,400,600}; learning_rate ∈[0.01,0.2] (log); num_leaves ∈{4,8,16,31}; min_child_samples ∈{10,20,40,80}; subsample ∈{0.7,0.85,1.0}; colsample_bytree ∈{0.5,0.7,0.9}; reg_lambda ∈[0.5,50] (log)					
Explainable boosting machine		learning_rate ∈[0.003,0.05] (log); interactions ∈{0,4,8,12}; max_bins ∈{128,256}; min_samples_leaf ∈{2, 4,8}; max_rounds ∈{1000,2000,4000}					

4.2. Statistical Assessment and Implementation

Differences between the proposed framework and each baseline were assessed on the 15 paired fold-level outcomes by the Wilcoxon signed-rank test. Because the training partitions of a repeated cross-validation overlap, the ordinary paired t statistic understates the variance; the corrected resampled statistic of Nadeau and Bengio [50] is therefore reported alongside it,

$$t = \frac{\bar{d}}{\sqrt{\left(\frac{1}{F} + \frac{n_{\text{test}}}{n_{\text{train}}}\right)\hat{\sigma}_d^2}}, \tag{18}$$

where $\bar{d}$ and $\hat{\sigma}_d^2$ are the mean and sample variance of the fold-level differences, $F = 15$ is the number of folds, and $n_{\text{train}} = 5634$ and $n_{\text{test}} = 1409$ are the partition sizes of the larger test folds; the ratio of test to training size differs by less than 0.0003 across folds.

All experiments were executed in Python version 3.16 on a single processor core. The penalised logistic regression and random forest were taken from scikit-learn, extreme gradient boosting from the XGBoost library, the light gradient boosting machine from LightGBM, and the explainable boosting machine from the InterpretML package [44]. The number of outer bags of the explainable boosting machine was fixed at two across every configuration, including the default, so that all configurations of that family carry the same computational weight. Tree-based Shapley values were obtained from the SHAP package [48]. The Arithmetic Optimization Algorithm was implemented from the description and reference implementation of its authors [46]. Random seeds were fixed and are recorded in Table 2.

5. Results and Discussion

The results are reported as follows: component tuning (Section 5.1), the learner families (Section 5.2), their statistical assessment (Section 5.3), the decision layer (Section 5.4), aggregation and search (Section 5.5), and attribution (Section 5.6). Section 5.7 compares the findings with previous studies and Section 5.8 gives the operational implications and the limitations.

5.1. The Contribution of Component Tuning

Table 3 compares each family under its library default against the two tuned selections, one made under the coupled objective of Equation (9) and one made under a ranking-only objective. The effect of tuning is large for the tree ensembles and modest for the additive models. Extreme gradient boosting rises from 0.8217 to 0.8483 in ranking quality and from 0.4308 to 0.4899 in the Matthews correlation coefficient, and the random forest rises from 0.8200 to 0.8470 and from 0.4346 to 0.4905. The explainable boosting machine, in contrast, moves only from 0.8485 to 0.8493, a change of 0.0008, which indicates that its library default already lies close to the best configuration in the search space. To keep the protocol identical across families, the configuration selected under the coupled objective is the one evaluated for all five in the tables that follow. The pattern is interpretable in terms of capacity: the untuned tree ensembles are permitted to grow deep enough to fit sampling noise in 5634 training records, whereas the additive families are constrained by construction and have little to gain from adjustment.

Table 3. Contribution of tuning. Each entry is the mean over 15 outer folds under the shared protocol, with the plateau-stabilised threshold applied to every configuration.

Family	AUC-ROC			MCC		
	Default	Coupled	Ranking	Default	Coupled	Ranking
Penalised logistic regression	0.8449	0.8482	**0.8487**	0.4762	**0.4884**	0.4872
Random forest	0.8200	0.8470	**0.8477**	0.4346	**0.4905**	0.4892
Extreme gradient boosting	0.8217	**0.8483**	0.8480	0.4308	0.4899	**0.4921**
Light gradient boosting machine	0.8360	0.8484	**0.8485**	0.4611	**0.4888**	**0.4888**
Explainable boosting machine	0.8485	**0.8493**	**0.8493**	0.4882	**0.4906**	**0.4906**

Bold formatting indicates the most significant values.

The coupled and the ranking-only objectives select configurations of indistinguishable quality. Extreme gradient boosting attains 0.4921 under the ranking objective against 0.4899 under the coupled objective, while penalised logistic regression attains 0.4884 against 0.4872 in the opposite direction. No family shows a consistent advantage for either. Coupling the search objective to the decision metric therefore does not change the quality of the configuration that is chosen, a result that anticipates the decomposition in Section 5.4: the information the decision layer adds is not information the tuner can exploit.

5.2. Comparison of the Learner Families

Table 4 reports the five criteria for every family and for the proposed hybrid, each averaged over the 15 outer folds with its standard deviation across folds. Every learner is far removed from the class-prior benchmark, which attains 0.7346 in accuracy purely by declining to predict the positive class and consequently scores zero on the F1 score and on the Matthews correlation coefficient. This contrast is worth stating explicitly, because accuracy values of the magnitude reported in Table 4 are often presented as strong outcomes when the trivial benchmark already reaches 0.7346 on this class distribution; the informative comparison is with the benchmark, not with zero.

The proposed hybrid obtains the highest mean value on three of the five criteria, namely 0.8499 in ranking quality, 0.6380 on the F1 score, and 0.4926 on the Matthews correlation coefficient, and shares the highest accuracy of 0.8057 with penalised logistic regression. On balanced accuracy the explainable boosting machine leads at 0.7682 against 0.7673 for the hybrid, a difference of 0.0009. The spread between the strongest model and the weakest family in ranking quality is 0.0029, from 0.8499 for the hybrid to 0.8470 for the random forest, which is smaller than the standard deviation across folds of any individual

family. On these 19 attributes the ranking layer is close to saturation, and the families that differ most in their inductive bias nonetheless converge to a narrow band.

Table 4. Predictive performance under repeated stratified five-fold cross-validation.

Model	AUC-ROC	F1	Accuracy	Balanced Accuracy	MCC
Penalised logistic regression	0.8482 ± 0.0091	0.6323 ± 0.0140	**0.8057** ± 0.0090	0.7662 ± 0.0108	0.4884 ± 0.0204
Random forest	0.8470 ± 0.0102	0.6328 ± 0.0143	0.8016 ± 0.0104	0.7671 ± 0.0110	0.4905 ± 0.0269
Extreme gradient boosting	0.8483 ± 0.0092	0.6350 ± 0.0135	0.8036 ± 0.0089	0.7679 ± 0.0110	0.4899 ± 0.0247
Light gradient boosting machine	0.8484 ± 0.0093	0.6358 ± 0.0149	0.8034 ± 0.0088	0.7661 ± 0.0107	0.4888 ± 0.0256
Explainable boosting machine	0.8493 ± 0.0107	0.6349 ± 0.0183	0.8046 ± 0.0081	**0.7682** ± 0.0130	0.4906 ± 0.0303
CATCH (proposed)	**0.8499** ± 0.0096	**0.6380** ± 0.0136	**0.8057** ± 0.0085	0.7673 ± 0.0106	**0.4926** ± 0.0277
Class-prior benchmark	0.5000 ± 0.0000	0.0000 ± 0.0000	0.7346 ± 0.0002	0.5000 ± 0.0000	0.0000 ± 0.0000

Bold formatting indicates the most significant values.

The mean receiver operating characteristic curves in Figure 3 display the same convergence: the six curves are visually superimposed over most of the range, separating slightly only in the region of low false-positive rate that matters most when a retention budget permits contacting a small fraction of the base. Figure 4 shows the fold-level dispersion behind the means of Table 4 and makes clear that between-fold variation exceeds between-model variation for every criterion, which is the condition under which a single-split comparison would be unreliable.

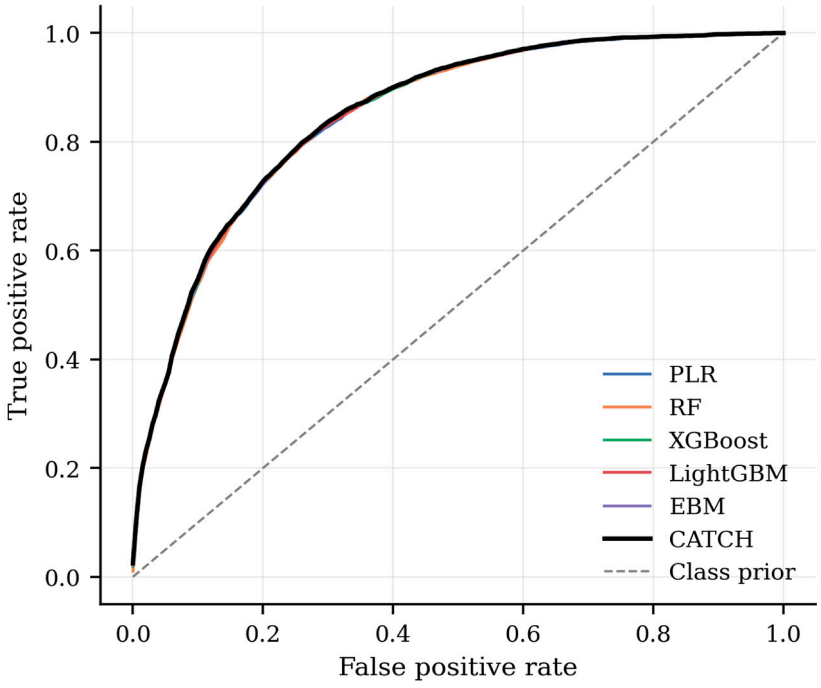
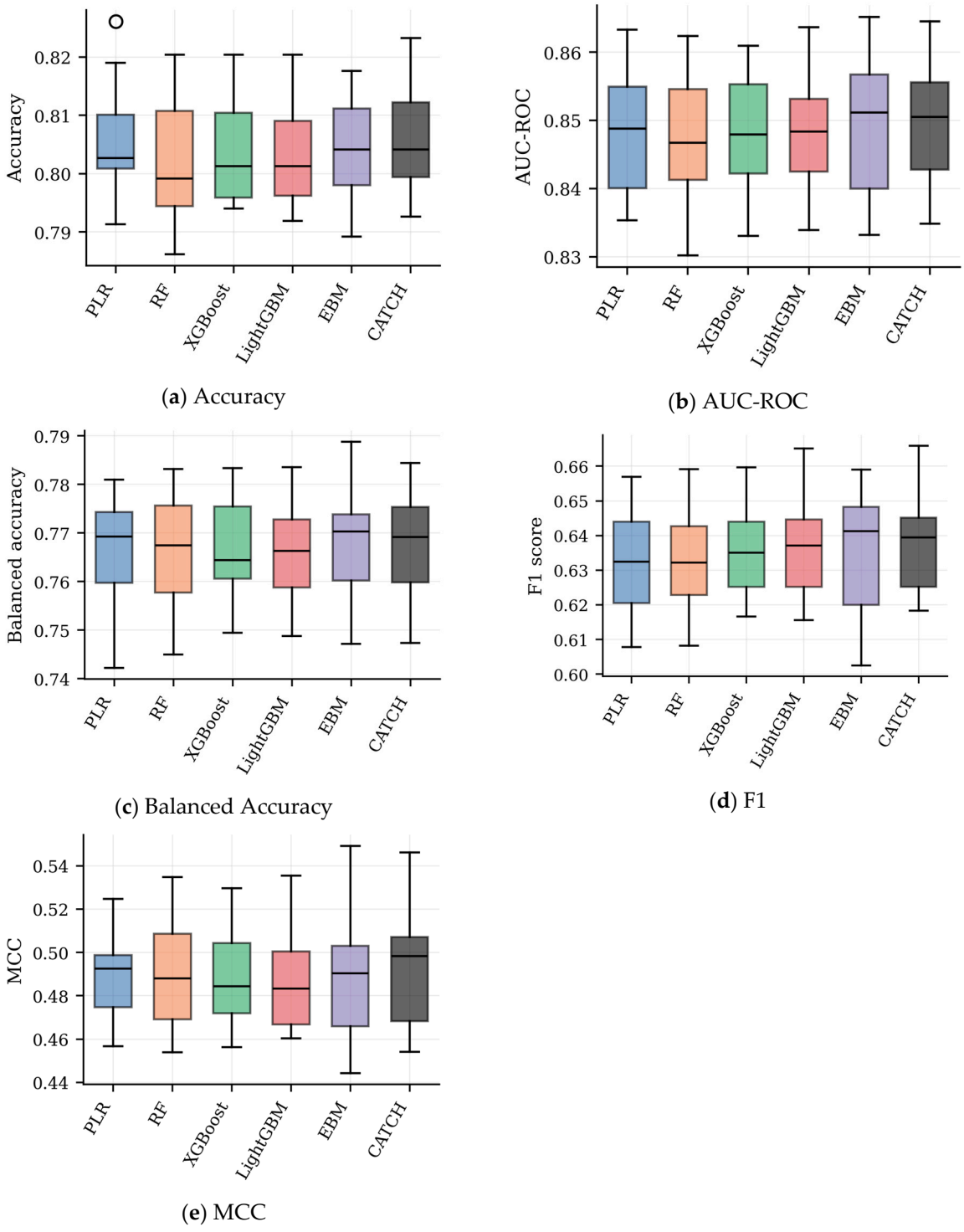


Figure 3. Mean receiver operating characteristic curves across the 15 outer folds.



(**a**) Accuracy (**b**) AUC-ROC

(**c**) Balanced Accuracy (**d**) F1

(**e**) MCC

Figure 4. Distribution across the 15 outer folds of each evaluation criterion.

5.3. Statistical Assessment

Table 5 pairs the hybrid against each baseline fold by fold. In ranking quality, the hybrid improves on four of the five baselines with signed-rank probabilities at or below 0.002, winning on 12 to 14 of the 15 folds, and the largest advantage is the 0.0028 recorded against the random forest. Against the explainable boosting machine, the advantage of 0.0005 is not distinguishable from zero, with a signed-rank probability of 0.330 and 8 winning folds out of 15. Under the conservative correction of Equation (18), which accounts for the overlap of training partitions, only the comparison with the random forest

retains a probability below 0.05, at 0.026. The signed-rank results are therefore evidence of a consistent direction.

Table 5. Paired comparison of the proposed hybrid against each baseline over the 15 outer folds. Positive differences favour the hybrid.

Criterion	Comparator	Mean Difference	Winning Folds	Signed-Rank p	Corrected p
AUC-ROC	Explainable boosting machine	+0.0005	8/15	0.330	0.579
AUC-ROC	Light gradient boosting machine	+0.0014	14/15	0.001 *	0.114
AUC-ROC	Penalised logistic regression	+0.0017	13/15	0.001 *	0.095
AUC-ROC	Random forest	+0.0028	14/15	0.000 *	0.026
AUC-ROC	Extreme gradient boosting	+0.0015	12/15	0.002 *	0.097
F1	Explainable boosting machine	+0.0031	9/15	0.303	0.590
F1	Light gradient boosting machine	+0.0022	10/15	0.083	0.382
F1	Penalised logistic regression	+0.0057	11/15	0.004 *	0.139
F1	Random forest	+0.0051	12/15	0.002 *	0.101
F1	Extreme gradient boosting	+0.0029	11/15	0.064	0.410
Accuracy	Explainable boosting machine	+0.0011	8/15	0.363	0.647
Accuracy	Light gradient boosting machine	+0.0023	11/15	0.005 *	0.147
Accuracy	Penalised logistic regression	−0.0000	7/15	0.977	1.000
Accuracy	Random forest	+0.0041	13/15	0.002 *	0.061
Accuracy	Extreme gradient boosting	+0.0020	11/15	0.047 *	0.272
Balanced accuracy	Explainable boosting machine	−0.0010	8/15	0.679	0.801
Balanced accuracy	Light gradient boosting machine	+0.0012	10/15	0.303	0.630
Balanced accuracy	Penalised logistic regression	+0.0010	8/15	0.489	0.692
Balanced accuracy	Random forest	+0.0002	7/15	0.847	0.953
Balanced accuracy	Extreme gradient boosting	−0.0006	5/15	0.454	0.808
MCC	Explainable boosting machine	+0.0020	8/15	0.762	0.754
MCC	Light gradient boosting machine	+0.0038	11/15	0.107	0.398
MCC	Penalised logistic regression	+0.0042	9/15	0.188	0.514
MCC	Random forest	+0.0021	6/15	0.639	0.702
MCC	Extreme gradient boosting	+0.0027	11/15	0.208	0.584

The same picture holds on the threshold-dependent criteria. The advantages of 0.0057 on the F1 score against penalised logistic regression and 0.0051 against the random forest carry signed-rank probabilities of 0.004 and 0.002, whereas every comparison on balanced accuracy is indistinguishable from zero, including the deficit of 0.0010 against the explainable boosting machine. The measured claim supported by Table 5 is therefore that the hybrid is never worse than any baseline on any criterion by a distinguishable margin, that it is distinguishably better than four baselines in ranking quality, and that it is equivalent to the explainable boosting machine. Given that the hybrid contains that model as one of its four members, equivalence rather than dominance is the outcome the construction would lead one to expect. In Table 5, an asterisk marks a signed-rank probability below 0.05. The final column reports the probability from the corrected resampled statistic of Equation (18).

5.4. The Contribution of the Decision Layer

Panel A of Table 6 averages the five families and the hybrid, each at the configuration selected under the coupled objective, over the three decision rules examined. Replacing the conventional fixed threshold of 0.5 with a threshold estimated on the inner partition raises the mean F1 score from 0.5834 to 0.6348, the mean balanced accuracy from 0.7126 to 0.7671, and the mean Matthews correlation coefficient from 0.4653 to 0.4901, while accuracy is essentially unchanged at 0.8041 against 0.8043. The plateau estimator of Equation (10) improves on the single maximiser on all three of the affected criteria, most clearly on the Matthews correlation coefficient, where it records 0.4901 against 0.4875.

Table 6. Contribution of the decision layer, evaluated at each family's configuration selected under the coupled objective.

Panel A: mean over the five families and the hybrid				
Decision rule	**F1**	**Accuracy**	**Balanced accuracy**	**MCC**
Fixed threshold 0.5	0.5834	**0.8043**	0.7126	0.4653
Inner maximiser	0.6343	0.8039	0.7665	0.4875
Plateau-stabilised (proposed)	**0.6348**	0.8041	**0.7671**	**0.4901**

Panel B: per family, fixed threshold against the plateau estimator						
Family	**F1**			**MCC**		
	Fixed 0.5	**Plateau**	**Difference**	**Fixed 0.5**	**Plateau**	**Difference**
Penalised logistic regression	0.5928	0.6323	+0.0394	0.4734	0.4884	+0.0150
Random forest	0.5670	0.6328	+0.0658	0.4531	0.4905	+0.0374
Extreme gradient boosting	0.5764	0.6350	+0.0586	0.4587	0.4899	+0.0312
Light gradient boosting machine	0.5865	0.6358	+0.0493	0.4671	0.4888	+0.0217
Explainable boosting machine	0.5890	0.6349	+0.0459	0.4687	0.4906	+0.0219
CATCH (proposed)	0.5889	0.6380	+0.0490	0.4709	0.4926	+0.0217

Panel B shows the same contrast family by family. The gain from estimating the threshold ranges from 0.0394 on the F1 score for penalised logistic regression to 0.0658 for the random forest, and on the Matthews correlation coefficient from 0.0150 to 0.0374. The ordering is informative: the families that benefit most are those whose default probability estimates are most compressed towards the centre of the interval, and the family that benefits least is the one whose output is already close to calibrated. For the proposed hybrid the improvement is 0.0490 on the F1 score and 0.0217 on the Matthews correlation coefficient.

Placing these figures next to Table 5 answers, for this dataset, the question posed in the introduction. The largest advantage the hybrid holds over any baseline in ranking quality is 0.0028, whereas moving from the fixed threshold to the estimated one is worth between 0.0150 and 0.0374 on the same Matthews correlation coefficient for which the between-model differences never exceed 0.0042. The decision layer is therefore worth several times more than the choice among learner families, and it costs nothing beyond an evaluation of an existing criterion curve on data already in hand.

Figure 5 traces the criterion curves that Equation (10) operates on. All three criteria are flat over a broad interior region and fall away sharply outside it, and for each the region of near-optimal values sits below the conventional cut-off marked on the plot. The curves also separate the criteria from one another: the plateau of balanced accuracy lies at lower thresholds than that of the F1 score, and that of the Matthews correlation coefficient at higher ones, which is why the framework estimates a separate threshold for each target criterion rather than a single operating point.

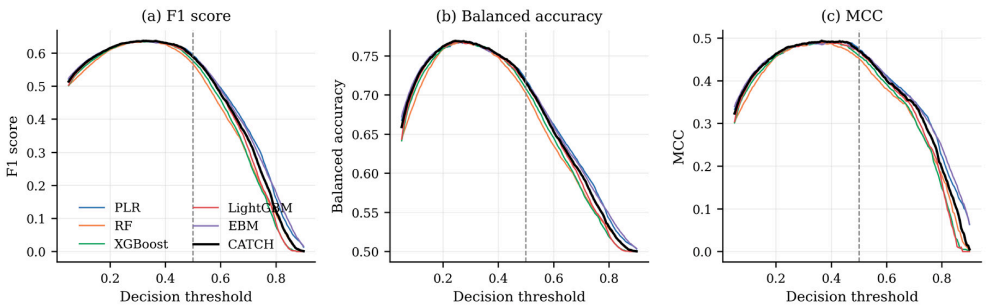


Figure 5. Criterion curves averaged over the 15 outer folds as a function of the decision threshold, for (**a**) the F1 score, (**b**) balanced accuracy, and (**c**) the Matthews correlation coefficient.

5.5. Aggregation Weighting and the Search Strategies

Panel A of Table 7 reports the verification of the Arithmetic Optimization Algorithm on five benchmark functions in 30 dimensions over 20 independent runs. The optimiser approaches the optimum on the separable unimodal function and on all three multimodal functions, with mean best values of 4.396×10^{-6} on Sphere, 1.645×10^{-6} on Rastrigin, 4.375×10^{-4} on Ackley, and 1.748×10^{-3} on Griewank, while stalling on the non-separable Rosenbrock valley at a mean of 2.792×10^{1}. The optimiser therefore converges well on four of the five functions and was accepted on that basis before being applied to the aggregation problem.

Table 7. Verification of the search machinery.

Panel A: benchmark verification of the reimplementation					
Function	**Class**	**Dimensions**	**Mean**	**Standard Deviation**	**Best**
Sphere	Unimodal	30	4.396×10^{-6}	1.803×10^{-6}	5.563×10^{-7}
Rosenbrock	Unimodal	30	2.792×10	2.444×10^{-1}	2.735×10
Rastrigin	Multimodal	30	1.645×10^{-6}	1.294×10^{-6}	3.553×10^{-14}
Ackley	Multimodal	30	4.375×10^{-4}	1.971×10^{-4}	1.543×10^{-5}
Griewank	Multimodal	30	1.748×10^{-3}	4.274×10^{-3}	9.448×10^{-6}
Panel B: aggregation weighting schemes					
Weighting scheme			**Inner objective *J***	**AUC-ROC**	**F1**	**MCC**
Uniform weights			0.6737	**0.8499**	**0.6380**	0.4926
Shrunk optimised weights			0.6737	**0.8499**	**0.6380**	0.4926
Random search			0.6756	0.8496	0.6349	0.4912
Arithmetic Optimization Algorithm			**0.6760**	0.8491	0.6351	0.4918
Uniform weights, five members			0.6738	**0.8499**	0.6369	**0.4933**

Panel B compares the four weighting schemes. The metaheuristic attains the highest value of the inner objective at 0.6760, ahead of random search at 0.6756 and of the uniform mixture at 0.6737, and yet on held-out data the ordering reverses: uniform weights give 0.8499 in ranking quality against 0.8496 for random search and 0.8491 for the metaheuristic, and 0.4926 on the Matthews correlation coefficient against 0.4912 and 0.4918. The shrinkage of Equation (8) reaches the same held-out values as the uniform mixture, because the one-standard-error rule drives the coefficient to its upper limit and returns the uniform point of the simplex. Figure 6 displays both halves of this outcome: panel (a) shows the metaheuristic climbing above random search on the inner criterion, and panel (b) places that climb beside the held-out result it produces.

The interpretation is that the four members are so highly correlated that the objective surface over the simplex is nearly flat, so additional optimisation effort purchases inner-sample fit rather than generalisation. This is the behaviour documented in the forecast-combination literature, where estimated optimal weights frequently fail to improve on equal weights because estimating the weights introduces bias and additional variance [29]. The search over aggregation weights therefore yields no gain over the uniform mixture, which the one-standard-error rule also returns, and the uniform mixture was carried forward as the operative form of Equation (6). The metaheuristic is reported here because the comparison is the evidence for that decision, not because it contributed to the final model. Panel B also records that admitting the random forest as a fifth member leaves ranking quality unchanged at 0.8499 and raises the Matthews correlation coefficient marginally to 0.4933, at the cost of the exact attribution property that Equation (16) requires; the four-member pool was retained for that reason.

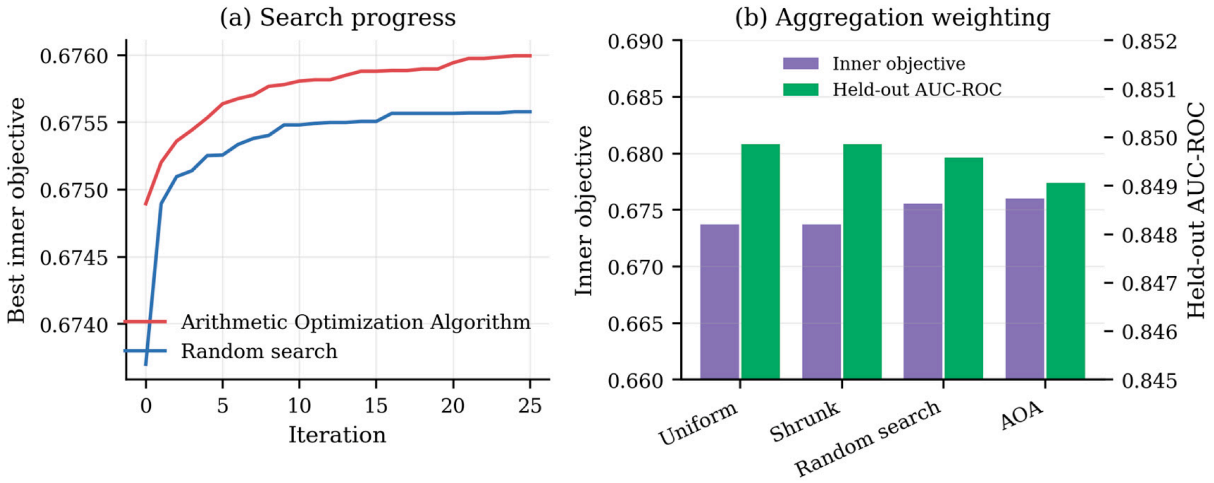


Figure 6. Aggregation search. (**a**) Best inner objective against iteration, averaged over the 15 outer folds, for the Arithmetic Optimization Algorithm and for random search under an equal evaluation budget. (**b**) Inner objective beside held-out ranking quality for each weighting scheme.

5.6. Attribution

Table 8 compares the two candidate attribution methods on the proposed hybrid across the 15 outer folds. Composed Shapley attribution is the more stable on every criterion examined: its mean pairwise Spearman rank correlation across folds is 0.952 against 0.850, its Kendall correlation 0.855 against 0.718, the mean Jaccard overlap of the five highest-ranked attributes 0.800 against 0.696, and the mean standard deviation of an attribute's rank 0.976 against 1.749, so that a rank obtained by permutation moves nearly twice as far between folds. The composed method was therefore selected, and the interpretation that follows rests on it alone. The two methods nonetheless agree closely in aggregate, with a Spearman correlation of 0.975 between their mean importance vectors, so the choice affects the precision of the ranking rather than its substance.

Table 8. Stability of the two attribution methods over the 15 outer folds.

Stability Criterion	Composed Shapley	Permutation Importance
Mean pairwise Spearman rank correlation	0.952	0.850
Mean pairwise Kendall rank correlation	0.855	0.718
Mean Jaccard overlap of the five highest-ranked attributes	0.800	0.696
Mean standard deviation of an attribute rank	0.976	1.749
Cross-method Spearman rank correlation		0.975

Table 9 ranks the 19 attributes by mean absolute Shapley value on the log-odds scale. Contract type dominates at 0.6423, followed by tenure at 0.4527 and internet service at 0.3491; the fourth-placed attribute, online security, contributes 0.1961, less than a third of the leader. Below the tenth position, the values fall under 0.1050, and for the lowest-ranked attributes the standard deviations across folds become comparable to the values themselves. Gender contributes 0.0099 and partner status 0.0101, both with standard deviations of 0.0088, which places them at the level of estimation noise; permutation importance assigns both a negative value of −0.00003, meaning that shuffling them left ranking quality marginally better than it was.

The two methods place thirteen of the nineteen attributes in identical positions and disagree by no more than three ranks anywhere. The largest disagreements concern technical support, ranked fifth by the composed method and eighth by permutation, and total charges, ranked eighth and fifth respectively. Both attributes are strongly associated with

others in the set, and permutation importance divides the shared contribution differently because permuting one attribute alone creates combinations that lie outside the observed data and forces the model to extrapolate, a known property of permutation-based importance under correlated attributes [51]. The disagreement is recorded here as a limitation on the precision of positions five to nine and is not pursued further. The ranking follows from the fitted models rather than from prior expectation, and two features of it deserve comment on that basis. First, the leading position of contract type is consistent with the descriptive spread in Table 1, where contract has the largest Cramér's V at 0.410 and a churn rate ranging from 2.8 to 42.7 per cent across its three levels. Second, the two monetary attributes rank eighth and ninth despite total charges carrying a point-biserial correlation of -0.198 and monthly charges $+0.193$ with the outcome; their marginal association is largely absorbed by tenure, with which total charges is mechanically linked, so the models extract little additional information from them once tenure is present. Figure 7 displays the importance values with their variation across folds and, in panel (b), the rank stability underlying Table 8.

Table 9. Attribute ranking under the selected attribution method.

Rank	Attribute	Mean $\lvert\phi_j\rvert$	Std. dev.	Permutation Value	Permutation Rank	Rank Difference
1	Contract	0.6423	0.0217	0.06067	1	+0
2	tenure	0.4527	0.0368	0.04841	2	+0
3	InternetService	0.3491	0.0304	0.02048	3	+0
4	OnlineSecurity	0.1961	0.0196	0.00465	4	+0
5	TechSupport	0.1524	0.0145	0.00271	8	−3
6	PaymentMethod	0.1417	0.0195	0.00286	6	+0
7	PaperlessBilling	0.1274	0.0196	0.00267	9	−2
8	TotalCharges	0.1174	0.0366	0.00389	5	+3
9	MonthlyCharges	0.1082	0.0231	0.00274	7	+2
10	MultipleLines	0.1050	0.0217	0.00194	10	+0
11	StreamingMovies	0.0930	0.0230	0.00149	11	+0
12	StreamingTV	0.0848	0.0141	0.00131	12	+0
13	OnlineBackup	0.0691	0.0118	0.00056	14	−1
14	SeniorCitizen	0.0442	0.0120	0.00061	13	+1
15	Dependents	0.0356	0.0121	0.00019	15	+0
16	DeviceProtection	0.0320	0.0030	0.00017	16	+0
17	PhoneService	0.0126	0.0043	0.00015	17	+0
18	Partner	0.0101	0.0088	−0.00003	18	+0
19	gender	0.0099	0.0088	−0.00003	19	+0

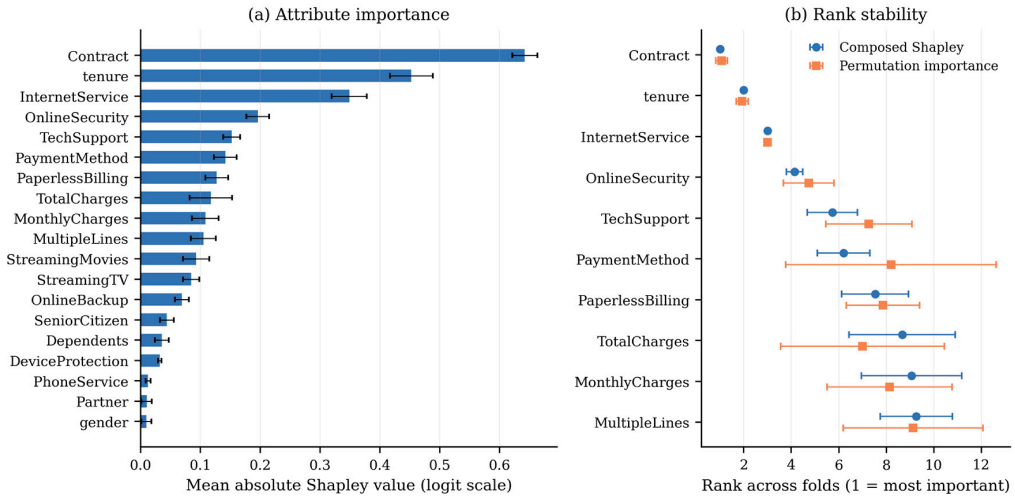


Figure 7. Attribution for the proposed hybrid. (**a**) Mean absolute Shapley value per attribute with the standard deviation across folds. (**b**) Rank of the ten leading attributes under each method, with the dispersion of that rank across folds.

5.7. Comparison with Previous Studies and Practical Significance

Three of the findings confirm earlier work. The narrow band into which the tuned models fall, with a spread of 0.0029 in ranking quality, agrees with Verbeke et al. [17], who found comparable performance across a large group of classifiers on eleven operator datasets, and with Coussement et al. [18] and Hand [20]. The use of Shapley explanations to identify the drivers of churn follows Poudel et al. [13], and the dominance of contract type and tenure in the attribution matches the marginal associations in Table 1. The size of the decision-layer effect, and the position of the near-optimal thresholds below 0.5 in Figure 5, agree with Zou et al. [9] and Lipton et al. [32], and the fact that the gain is obtained without resampling is consistent with [17,34].

Two findings differ from part of the literature. First, the clear advantage reported for the hybrid sequence model reviewed in Section 2 is not observed here: once every constituent is tuned under the same budget and decision rule, the margin of the aggregate over its best member contracts to 0.0005 in ranking quality, in line with El Attar and El-Hajj [6], whose soft-voting ensemble did not exceed its strongest member, and with the forecast combination puzzle [29]. Second, the area under the curve of 0.8499 obtained here is below the 0.932 reported in [6] on the same records. The two values are not directly comparable, because the latter was obtained with engineered features and synthetic minority oversampling, whereas the present protocol leaves the attributes and the class distribution untouched within a nested design [35]; the comparison illustrates how strongly a reported level depends on the protocol.

The practical significance of the results is correspondingly uneven. Differences of the order of 0.002 between tuned learners will not change a contact list in any material way. The decision-layer effect is of a different order: for the hybrid, replacing the fixed threshold raises the F1 score from 0.5889 to 0.6380 and the Matthews correlation coefficient from 0.4709 to 0.4926 at no additional modelling cost (Table 6). Expressing that improvement in monetary terms requires the cost of a contact, the value of a retained subscriber and the acceptance rate of an offer, none of which is recorded in the public dataset, so the evaluation is confined to statistical criteria. The framework is nevertheless compatible with value-based evaluation: the threshold estimator of Equation (10) accepts any criterion curve, so maximum-profit measures [17,36] can serve as its target, a campaign budget corresponds to choosing the threshold by contact fraction, and uplift modelling [37,38] requires data from retention campaigns with treated and control groups.

5.8. Operational Implications and Limitations

Three recommendations follow from the results, and each rests on the selected attribution method and on the tabulated outcomes rather than on external expectation. They are of two kinds, which are kept apart here: the first describes an association observed in this particular dataset, whereas the second and third concern procedure and do not depend on the dataset.

The first concerns where retention effort is directed. Contract type, tenure, and internet service jointly account for just over half of the summed attribution in Table 9, at 0.6423, 0.4527 and 0.3491 against 0.1961 for the next attribute. Since contract type is the one among the three that an operator can alter directly through an offer, migration of month-to-month subscribers onto term agreements is the intervention with the largest support in these results. This is an association in the records of one operator, not an estimated effect: the attribution identifies the subscribers the models regard as at risk, not those who would respond to an offer [37,38]. It is therefore best treated as a hypothesis to be tested in a controlled campaign, and not as a general managerial rule. Tenure supplies the timing rather than the instrument: because attributed importance is concentrated there and the

descriptive relation in Figure 2 declines steeply over the first year, the same offer is worth more early in the relationship than later.

The second concerns the operating point. Table 6 shows that choosing the threshold on training data is worth more than choosing among learner families, so an operator adopting any of the five baselines gains more by estimating the cut-off than by replacing the model. Unlike the first, this recommendation is procedural and can be adopted on any dataset, although the size of the gain will vary with prevalence and calibration. The threshold must be estimated per objective: the plateau of balanced accuracy sits below that of the F1 score, and that of the Matthews correlation coefficient above it, in Figure 5, so campaigns that pursue different objectives call for different cut-offs from the same scores. The third concerns the scores themselves. Figure 8 shows that the predicted probabilities track observed churn frequencies closely across the range for the additive members and the hybrid, and deviate more for the random forest, most visibly for predicted probabilities between 0.1 and 0.2 and again at the upper end. Where a retention budget is allocated in proportion to expected loss rather than by a fixed contact count, that difference matters more than the ranking differences in Table 4, since the allocation depends on the value of the probability and not only on its order. On this dataset, two attributes should not enter targeting at all. Gender and partner status sit at the bottom of Table 9 at 0.0099 and 0.0101, within the noise of the estimate, and using either to select subscribers for an offer would introduce a demographic criterion the models do not support.

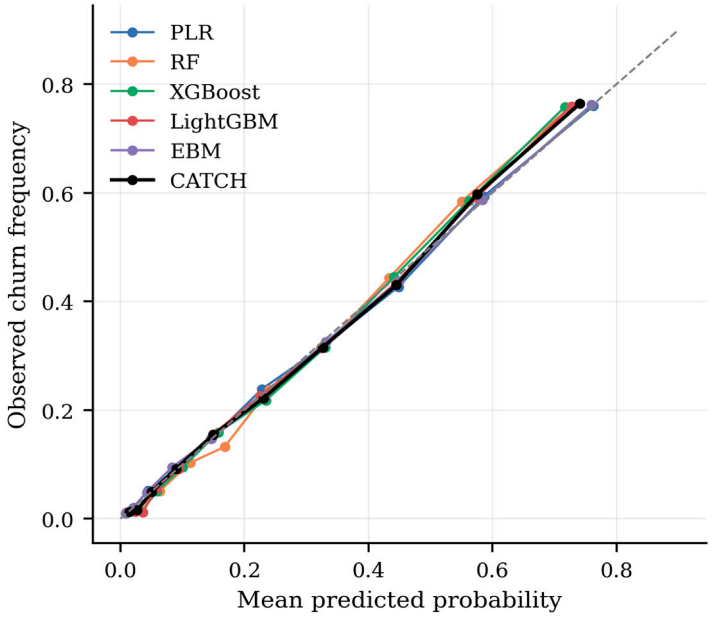


Figure 8. Reliability of the predicted probabilities, averaged over the 15 outer folds using ten quantile bins.

Four limitations bound the conclusions. The evaluation rests on a single public dataset of 7043 records from one operator, so the saturation of the ranking layer observed here need not transfer to settings with richer behavioural attributes, such as the call-graph descriptors that raised the area under the curve from 84 to 93.3 per cent in the study of Ahmad et al. [4]. The dataset was chosen because closely related studies report on it or on closely similar public telecommunications data [6,13], which makes the protocol comparable. A measure of external validity is supported by the fact that the protocol and the decision layer use no dataset-specific information and that the qualitative findings, namely the convergence of tuned learners, the sufficiency of threshold adjustment without resampling and the failure of optimised combining weights to generalise, have each been reported independently

on other data or explained theoretically [17,29,34]. The records carry no time index, so the protocol cannot examine how the models age between the moment of fitting and the moment of deployment. The comparison is confined to the 19 native attributes and excludes derived quantities, which keeps the contrast between families clean but leaves open whether construction of new attributes would lift the band in which they all sit. Finally, the number of outer folds is 15, and the corrected statistic of Equation (18) shows that this resolution distinguishes differences of roughly 0.003 in ranking quality but not the smaller margins that separate the leading models in Table 4. Beyond these, the evaluation uses statistical criteria only, so that costs, budgets, profit and treatment effects lie outside its scope; and the attribute-level findings are specific to the dataset analysed.

6. Conclusions

This study evaluated five learner families and a proposed Composable Additive-Tree Convex Hybrid on a public benchmark of 7043 subscriber records under a single nested protocol in which the resampling design, the search budget, the search objective and the decision rule were held identical for every candidate. The hybrid combines four learners whose log-odds admit an exact additive decomposition, so that its attribution follows exactly from those of its members, and pairs them with a threshold estimated from the near-optimal plateau of the target criterion on the training partition.

The hybrid recorded the best mean value on ranking quality at 0.8499, on the F1 score at 0.6380, and on the Matthews correlation coefficient at 0.4926, and shared the best accuracy at 0.8057. Signed-rank tests placed it above four of the five baselines in ranking quality and on a level with the explainable boosting machine, which it contains. These margins are small; the hybrid is accordingly offered as an exactly attributable aggregate that is never distinguishably worse than its members. All six models stand far above the class-prior benchmark, whose accuracy of 0.7346 is obtained without predicting a single churn event.

The decomposition of these gains is the more consequential result. Tuning the tree ensembles was worth up to 0.0270 in ranking quality, estimating the decision threshold rather than fixing it at 0.5 was worth between 0.0150 and 0.0374 on the Matthews correlation coefficient, and aggregating four learners was worth 0.0005 in ranking quality over the best single member. Optimising the aggregation weights improved the inner objective while lowering held-out performance, and uniform weights were carried forward on that evidence. On this benchmark, therefore, the remaining gains sit in the tuning and decision layers rather than in the choice of hypothesis class, and a practitioner is better served by estimating the operating point than by exchanging one tabular learner for another. Composed Shapley attribution proved substantially more stable than permutation importance across folds and identified contract type, tenure, and internet service as the dominant drivers, with gender and partner status at the level of estimation noise. The methodological contribution therefore rests on two components rather than on a new learner: the decision layer, which improved on the fixed threshold on the F1 score and the Matthews correlation coefficient for every model, including the hybrid, and an aggregation whose attribution is exact and more stable across folds than permutation importance.

The directions for future work are as follows. The first step is replication of the protocol on further operator datasets and in other subscription sectors; the ordering of the contributions, with the decision layer ahead of tuning and tuning ahead of aggregation, is expected to persist wherever the positive class is a minority, whereas the contribution of aggregation may grow where the learner families disagree more. The second is the addition of behavioural, network and temporal attributes [4,11], with which the ranking layer is expected to leave the saturation observed on 19 static attributes. The third is evaluation under temporal partitions, where the plateau estimator is expected to transfer more reliably

than the single maximiser. The fourth is a profit-based target for the decision layer and the comparison of risk-based with uplift-based targeting on data from randomised campaigns.

Author Contributions: A.B.: conceptualization, methodology and validation; O.R.A.: methodology, formal analysis and draft preparation; A.B.A.: writing, original draft preparation and supervision. All authors have read and agreed to the published version of the manuscript.

Funding: The authors declare that no funding was received.

Data Availability Statement: The data obtained through the experiments are available upon request from the corresponding author.

Conflicts of Interest: The authors declare that there are no conflicts of interest.

References

1. Wagh, S.K.; Andhale, A.A.; Wagh, K.S.; Pansare, J.R.; Ambadekar, S.P.; Gawande, S.H. Customer churn prediction in telecom sector using machine learning techniques. *Results Control Optim.* **2024**, *14*, 100342. [CrossRef]
2. Guliyev, H.; Yerdelen Tatoğlu, F. Customer churn analysis in banking sector: Evidence from explainable machine learning models. *J. Appl. Microeconom.* **2021**, *1*, 85–99. [CrossRef]
3. Lalwani, P.; Mishra, M.K.; Chadha, J.S.; Sethi, P. Customer churn prediction system: A machine learning approach. *Computing* **2022**, *104*, 271–294. [CrossRef]
4. Ahmad, A.K.; Jafar, A.; Aljoumaa, K. Customer churn prediction in telecom using machine learning in big data platform. *J. Big Data* **2019**, *6*, 28. [CrossRef]
5. Khattak, A.; Mehak, Z.; Ahmad, H.; Asghar, M.U.; Asghar, M.Z.; Khan, A. Customer churn prediction using composite deep learning technique. *Sci. Rep.* **2023**, *13*, 17294. [CrossRef] [PubMed]
6. El Attar, A.; El-Hajj, M. Explainable AI-driven customer churn prediction: A multi-model ensemble approach with SHAP-based feature analysis. *Front. Artif. Intell.* **2026**, *9*, 1748799. [CrossRef] [PubMed]
7. Manzoor, A.; Atif Qureshi, M.; Kidney, E.; Longo, L. A Review on Machine Learning Methods for Customer Churn Prediction and Recommendations for Business Practitioners. *IEEE Access* **2024**, *12*, 70434–70463. [CrossRef]
8. Imani, M.; Joudaki, M.; Beikmohammadi, A.; Arabnia, H.R. Customer Churn Prediction: A Systematic Review of Recent Advances, Trends, and Challenges in Machine Learning and Deep Learning. *Mach. Learn. Knowl. Extr.* **2025**, *7*, 105. [CrossRef]
9. Zou, Q.; Xie, S.; Lin, Z.; Wu, M.; Ju, Y. Finding the Best Classification Threshold in Imbalanced Classification. *Big Data Res.* **2016**, *5*, 2–8. [CrossRef]
10. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6. [CrossRef] [PubMed]
11. Óskarsdóttir, M.; Bravo, C.; Verbeke, W.; Sarraute, C.; Baesens, B.; Vanthienen, J. Social network analytics for churn prediction in telco: Model building, evaluation and network architecture. *Expert Syst. Appl.* **2017**, *85*, 204–220. [CrossRef]
12. Chang, V.; Hall, K.; Xu, Q.A.; Amao, F.O.; Ganatra, M.A.; Benson, V. Prediction of Customer Churn Behavior in the Telecommunication Industry Using Machine Learning Models. *Algorithms* **2024**, *17*, 231. [CrossRef]
13. Poudel, S.S.; Pokharel, S.; Timilsina, M. Explaining customer churn prediction in telecom industry using tabular machine learning models. *Mach. Learn. Appl.* **2024**, *17*, 100567. [CrossRef]
14. Asif, D.; Arif, M.S.; Mukheimer, A. A data-driven approach with explainable artificial intelligence for customer churn prediction in the telecommunications industry. *Results Eng.* **2025**, *26*, 104629. [CrossRef]
15. Christopher, D.; Anand, G. Comparative Analysis of Predictive Models for Customer Churn Prediction in the Telecommunication Industry. In Proceedings of the 2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE), Gautam Buddha Nagar, India, 9–11 May 2024; pp. 534–539. [CrossRef]
16. Hao, M. Telecom Customer Churn Prediction Based on BiGRU-Attention-XGBoost Model. In Proceedings of the 2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT), Nanjing, China, 29–31 March 2024; pp. 1050–1054. [CrossRef]
17. Verbeke, W.; Dejaeger, K.; Martens, D.; Hur, J.; Baesens, B. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. *Eur. J. Oper. Res.* **2012**, *218*, 211–229. [CrossRef]
18. Coussement, K.; Lessmann, S.; Verstraeten, G. A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry. *Decis. Support Syst.* **2017**, *95*, 27–36. [CrossRef]
19. Geiler, L.; Affeldt, S.; Nadif, M. A survey on machine learning methods for churn prediction. *Int. J. Data Sci. Anal.* **2022**, *14*, 217–242. [CrossRef]
20. Hand, D.J. Classifier Technology and the Illusion of Progress. *Stat. Sci.* **2006**, *21*, 1–14. [CrossRef]

21. Grinsztajn, L.; Oyallon, E.; Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: New York, NY, USA, 2022; pp. 507–520. [CrossRef]
22. Shwartz-Ziv, R.; Armon, A. Tabular data: Deep learning is not all you need. *Inf. Fusion* **2022**, *81*, 84–90. [CrossRef]
23. Probst, P.; Boulesteix, A.-L.; Bischl, B. Tunability: Importance of Hyperparameters of Machine Learning Algorithms. *J. Mach. Learn. Res.* **2019**, *20*, 1–32.
24. De Bock, K.W.; Van den Poel, D. Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models. *Expert Syst. Appl.* **2012**, *39*, 6816–6826. [CrossRef]
25. De Caigny, A.; Coussement, K.; De Bock, K.W. A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. *Eur. J. Oper. Res.* **2018**, *269*, 760–772. [CrossRef]
26. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef] [PubMed]
27. Wolpert, D.H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]
28. Breiman, L. Stacked regressions. *Mach. Learn.* **1996**, *24*, 49–64. [CrossRef]
29. Claeskens, G.; Magnus, J.R.; Vasnev, A.L.; Wang, W. The forecast combination puzzle: A simple theoretical explanation. *Int. J. Forecast.* **2016**, *32*, 754–762. [CrossRef]
30. Chicco, D.; Jurman, G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Min.* **2023**, *16*, 4. [CrossRef] [PubMed]
31. Holzmann, H.; Klar, B. Robust performance metrics for imbalanced classification problems. *arXiv* **2026**, arXiv:2404.07661. [CrossRef]
32. Lipton, Z.C.; Elkan, C.; Naryanaswamy, B. Optimal Thresholding of Classifiers to Maximize F1 Measure. In *Machine Learning and Knowledge Discovery in Databases*; Calders, T., Esposito, F., Hüllermeier, E., Meo, R., Eds.; Springer: Berlin/Heidelberg, Germany, 2014; pp. 225–239. [CrossRef] [PubMed]
33. Elkan, C. The Foundations of Cost-Sensitive Learning. In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, Seattle, WA, USA, 4–10 August 2001.
34. van den Goorbergh, R.; van Smeden, M.; Timmerman, D.; Van Calster, B. The harm of class imbalance corrections for risk prediction models: Illustration and simulation using logistic regression. *J. Am. Med. Inform. Assoc.* **2022**, *29*, 1525–1534. [CrossRef] [PubMed]
35. Cawley, G.C.; Talbot, N.L.C. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *J. Mach. Learn. Res.* **2010**, *11*, 2079–2107.
36. Verbraken, T.; Verbeke, W.; Baesens, B. A Novel Profit Maximizing Metric for Measuring Classification Performance of Customer Churn Prediction Models. *IEEE Trans. Knowl. Data Eng.* **2013**, *25*, 961–973. [CrossRef]
37. Ascarza, E. Retention Futility: Targeting High-Risk Customers Might be Ineffective. *J. Mark. Res.* **2018**, *55*, 80–98. [CrossRef]
38. Devriendt, F.; Berrevoets, J.; Verbeke, W. Why you should stop predicting customer churn and start using uplift models. *Inf. Sci.* **2021**, *548*, 497–515. [CrossRef]
39. CML_AMP_Churn_Prediction/raw/WA_Fn-UseC_-Telco-Customer-Churn-.csv at master · cloudera/CML_AMP_Churn_ Prediction. GitHub. Available online: https://github.com/cloudera/CML_AMP_Churn_Prediction/blob/master/raw/WA_Fn-UseC_-Telco-Customer-Churn-.csv (accessed on 9 September 2026).
40. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
41. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; in KDD '16; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794. [CrossRef]
42. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: New York, NY, USA, 2017. Available online: https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html (accessed on 14 May 2025).
43. Lou, Y.; Caruana, R.; Gehrke, J.; Hooker, G. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; in KDD '13; Association for Computing Machinery: New York, NY, USA, 2013; pp. 623–631. [CrossRef]
44. Nori, H.; Jenkins, S.; Koch, P.; Caruana, R. InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv* **2019**, arXiv:1909.09223. [CrossRef]
45. Bergstra, J.; Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn Res.* **2012**, *13*, 281–305.
46. Abualigah, L.; Diabat, A.; Mirjalili, S.; Abd Elaziz, M.; Gandomi, A.H. The Arithmetic Optimization Algorithm. *Comput. Methods Appl. Mech. Eng.* **2021**, *376*, 113609. [CrossRef]
47. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*; in NIPS'17; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 4768–4777.

48. Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 56–67. [CrossRef] [PubMed]
49. Fisher, A.; Rudin, C.; Dominici, F. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *J. Mach. Learn. Res.* **2019**, *20*, 1–81.
50. Nadeau, C.; Bengio, Y. Inference for the Generalization Error. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 1999. Available online: https://proceedings.neurips.cc/paper_files/paper/1999/hash/7d12b66d3df6af8d429c1a357d8b9e1a-Abstract.html (accessed on 10 September 2026).
51. Hooker, G.; Mentch, L.; Zhou, S. Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance. *Stat. Comput.* **2021**, *31*, 82. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.