Article

Predicting Risk Profiles of New Customers in the Insurance Sector with Machine Learning Models

Finn L. Solly [1], Raquel Soriano-Gonzalez [2] and Angel A. Juan [2,3]*

[1] Esade Business School, Ramon Llull University, 59 Torre Blanca Av., 08172 Sant Cugat, Spain; finnluca.solly@alumni.esade.edu (F.L.S.)
[2] CIGIP - ValgrAI, Universitat Politècnica de València, Ferrandiz-Carbonell, 03802 Alcoy, Spain; rsorgon@epsa.upv.es (R.S.)
[3] Euncet Business School, Universitat Politècnica de Catalunya, Mas Rubial, 08225, Terrassa, Spain
* Correspondence: ajuanp@upv.es (A.A.J.)

Abstract: Classifying new customers in the insurance industry is challenging due to the absence of historical data, strong class imbalance, heavy-tailed loss distributions, and strict operational constraints. Standard machine learning models often fail to flag high-risk clients because they optimize for overall accuracy rather than business-critical outcomes. We propose a business-aware classification methodology based on a balanced bagging ensemble model. Our methodology is designed to account for the asymmetric cost structure and customer selection limits that exist in real-world insurance settings. It preserves the minority class while systematically adjusting the majority sampling ratio, guided by a profit-based objective function constrained by a maximum threshold for customer omission. Using a large-scale auto insurance dataset, our method improves the identification of loss-generating customers and increases net business value while maintaining regulatory and market compliance. This study contributes to business-optimized machine learning and provides a practical approach for risk-driven decision-making in the insurance sector.

Keywords: insurance risk modeling; high-risk customer detection; profit-aware machine learning.

Citation: Solly, F.L.; Soriano-Gonzalez, R.; Juan, A.A. Predicting Risk Profiles of New Customers in the Insurance Sector with Machine Learning Models. *Journal Not Specified* **2024**, *1*, 0. https://doi.org/

Received:
Revised:
Accepted:
Published:

1. Introduction

Accurate customer risk assessment is essential in the insurance industry to maintain profitability and competitiveness. Misclassification of customers can cause considerable financial losses, as poor risk assessment often leads to underpriced premiums and rising claims costs [1]. Predicting the performance of new customers is particularly difficult, as historical data is scarce and the costs of misclassification are high [2]. Customer classification is further complicated by severe class imbalance. Loss-generating customers typically account for only 10% to 20% of the portfolio but are responsible for 60% to 80% of total losses [3]. Standard binary classification approaches tend to favor the majority class, resulting in poor identification of high-loss customers [4]. Business constraints add to the challenge, since insurers cannot omit large numbers of customers without risking business volume and market share [5].

The financial implications of misclassification are critical. False negatives (misclassifying high-risk customers as low-risk) lead to direct financial losses through premiums that fail to cover claims. False positives (rejecting profitable customers) create opportunity costs [6]. This trade-off requires methods that balance risk identification with business sustainability. Most existing work addresses these issues through feature selection and model optimization. Soriano-Gonzalez et al. [7], for example, predicted new customer performance using XGBoost and LightGBM, combined with extensive feature engineering and profitability thresholds. However, their method does not directly address class imbalance or asymmetric misclassification costs [8]. It relies on arbitrary economic thresholds and does not incorporate operational constraints such as acceptable rejection rates or business risk tolerance. Nor does it explore ensemble strategies tailored to imbalance or apply business-driven principles such as the Pareto rule for threshold selection [9].

This paper develops a classification framework that combines a balanced ensemble approach with business-aware thresholding to address these limitations. We use the Pareto principle to determine thresholds for high-loss classification and train a balanced ensemble of XGBoost models on randomly sampled subsets of the data. Grid search is applied to optimize sampling and ensemble parameters. We benchmark our approach against the method in Soriano-Gonzalez et al. [7] to assess the effectiveness of balanced ensembles compared with traditional supervised classifiers. Figure 1 shows a step-by-step breakdown of this proposed methodology. Our approach tackles the key challenges of predicting performance of new customers (for which no historical data exists) under class imbalance and heavy-tailed loss distributions.

The remainder of the paper is organized as follows. Section 2 reviews related work. Our methodology involves three steps: (i) preparing and analyzing the data, which is described in Section 3; (ii) defining class imbalance within business constraints. as introduced in Section 4; and (iii) developing a balanced ensemble classifier, which is discussed in Section 5. Section 6 describes the sampling strategy, while Section 7 presents the results from multiple tests and datasets. Section 8 discusses the economic impact for the company of misclassified customers. Section 9 provides a comparison of our approach against the base benchmark using real-life data, while Section 10 performs the same comparison but this time using publicly available data, which has been synthetically generated. Finally, Section 11 summarizes the conclusions and outlines directions for future work.

2. Related Work

Machine learning methods are increasingly used in the insurance sector to support risk assessment, underwriting, and claims management. Their ability to process large datasets and support tasks such as classification and prediction has been well documented in various financial applications [10–12]. In life and auto insurance, models such as logistic regression, decision trees, random forest, support vector machines, naive Bayes, and XGBoost have been used to predict customer risk levels and claim outcomes [13,14]. These models reduce manual work in underwriting and improve operational efficiency. In the context of auto insurance, studies have shown that tree-based ensemble methods outperform traditional general linear models, particularly for predicting claim frequency and severity in imbalanced datasets [15]. Ensemble methods such as bagging, boosting, and stacking have proven effective in improving classification accuracy across various insurance applications, including fraud detection and high-cost claim prediction [16–18]. Techniques such as XGBoost and explainable boosting machines have been applied to maintain both high accuracy and model transparency [19]. For instance, Soriano-Gonzalez et al. [7] proposed a profit-based objective to address asymmetric misclassification costs in insurance classification. However, their approach had several limitations. They did not apply techniques to address class imbalance, lacked control over customer omission rates, and used fixed thresholds without systematic tuning. Their model was also limited to a single boosting algorithm, reducing its flexibility and robustness.

Several studies have focused on imbalanced data problems, as events like fraud or severe claims are rare compared to the broader customer base [20]. To address this, researchers have applied undersampling, oversampling (e.g., SMOTE), class weighting, and adaptive ensemble approaches. Baran and Rola [21] and Khamesian et al. [22] combined decision trees or neural networks with resampling techniques to improve the detection of rare insurance events. Similar challenges are present in medical applications, where rare event prediction has high ethical and operational implications [23]. Nonetheless, many machine learning studies in insurance still optimize only for statistical performance measures such as AUC or F1-score, without incorporating business constraints like asymmetric costs or acceptance limits. For example, thresholds are often set at $p = 0.5$ with no justification based on cost, regulation, or operational targets. Some approaches from other domains,

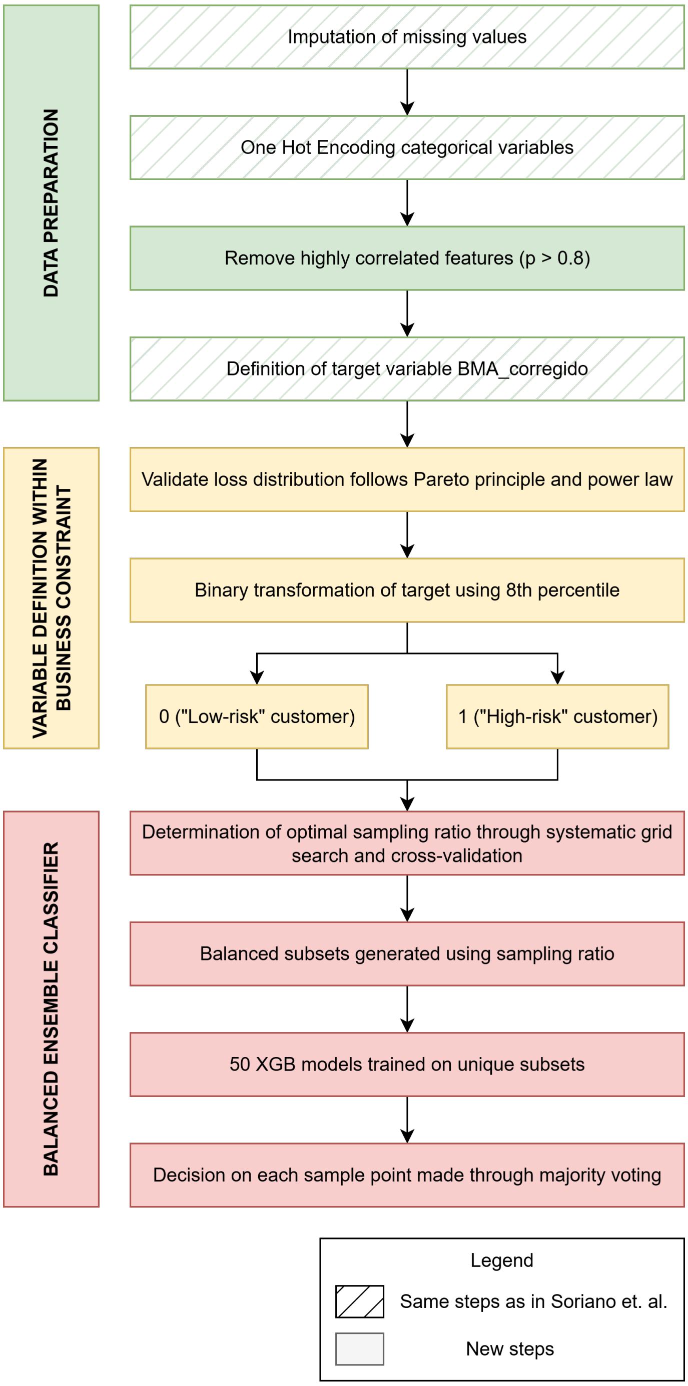
DATA PREPARATION
Imputation of missing values
One Hot Encoding categorical variables
Remove highly correlated features (p > 0.8)
Definition of target variable BMA_corregido
VARIABLE DEFINITION WITHIN BUSINESS CONSTRAINT
Validate loss distribution follows Pareto principle and power law
Binary transformation of target using 8th percentile
0 ("Low-risk" customer)
1 ("High-risk" customer)
BALANCED ENSEMBLE CLASSIFIER
Determination of optimal sampling ratio through systematic grid search and cross-validation
Balanced subsets generated using sampling ratio
50 XGB models trained on unique subsets
Decision on each sample point made through majority voting
Legend
Same steps as in Soriano et. al.
New steps

Figure 1. Methodology workflow of the proposed model.

such as online recommendation systems, have applied reinforcement learning and multitask neural networks to optimize conversion rates [24], but these are mostly limited to digital channels. In contrast, profit-based objectives in insurance remain underexplored, although some work has looked into economic performance in lapse risk modeling [25]. Our study builds on this gap by introducing a balanced bagging ensemble that includes multiple classifiers to manage severe class imbalance while preserving the distribution of rare classes. We integrate an explicit constraint to cap customer omission at 8%, ensuring operational feasibility. The sampling ratio (r) is optimized using stratified cross-validation and sensitivity analysis, and majority voting is applied to improve model robustness and generalization.

Beyond risk prediction, ML has also supported tasks such as customer segmentation. Clustering techniques including k-means, hierarchical clustering, and DBSCAN have been used to group customers with similar characteristics, supporting targeted marketing and service strategies [26]. However, classification of new or low-data customers remains limited, especially when data acquisition is costly or incomplete [27]. In such settings, the financial impact of misclassification can be high, but few studies provide solutions to handle this effectively [28]. Recent developments in deep learning have further expanded machine learning applications in insurance. Convolutional and recurrent neural networks, including long short-term memory variants, have been used for processing unstructured data, such as images of damaged vehicles or sequential customer records [29–31]. Natural language processing has also been integrated with machine learning to analyze text from policy documents, claims, and customer interactions [32–34]. Despite these advances, model interpretability remains a challenge, especially for deep learning. Insurers need models that are not only accurate but also explainable to meet regulatory requirements and maintain user trust [35,36]. Tools such as SHAP (Shapley additive explanations) and LIME (local interpretable model-agnostic explanation) have been adopted to improve transparency of model decisions [37,38]. Another ongoing concern is data privacy and compliance with regulations such as the General Data Protection Regulation in the European Union [39,40]. These constraints must be considered in the practical deployment of machine learning models in insurance workflows.

3. Data Preparation and Statistical Foundation

The empirical analysis uses a real-world insurance dataset with 116, 934 customer records covering 2016 to 2023. The initial dataset included 196 features on demographics, vehicle attributes, policy details, and historical claims. To ensure applicability to prospective customers, we removed features unavailable at acquisition, and other policy-specific variables. Following Soriano-Gonzalez et al. [7], we applied standard pre-processing. Unicode normalization and column name standardization ensured computational compatibility. Missing values were imputed using SoftImputer, an iterative matrix completion method. Customers lacking SINCO data (i.e., data from the Information System of the Insurance Compensation Consortium in Spain) were omitted, leaving 51, 618 observations with 196 features.

To address multi-collinearity, we computed Pearson correlation coefficients. For feature pairs with correlation > 0.8, we chosen the feature more correlated with the target, customer's average annual benefit. This removed 54 variables, yielding 142 features while preserving predictive information. The target variable poses major challenges for standard machine learning. The Jarque–Bera test [41] (1.81×10^8, $p < 0.001$) rejects normality, and Anderson–Darling tests [42] reject normal, exponential, logistic, and Gumbel probability distributions at all significance levels. Losses show heavy-tailed behavior, concentrating business risk. Of 51, 618 customers, 9, 984 (19.3%) generated negative returns, with total losses of 11, 328, 603 euros, while the portfolio produced only 909, 881 euros in net benefit before intervention. Thus, loss-generating customers impose losses more than 12 times the net benefit, underscoring the need for accurate identification.

Following Clauset et al. [43], we conducted a power-law analysis of tail behavior. Maximum likelihood estimation gave an exponent $\alpha = 2.381 \pm 0.057$ with $x_{\min} = 3,454$. The Kolmogorov–Smirnov test ($D = 0.0197$, with $p = 0.652$) indicated good fit. Similarly, semi-parametric bootstrap tests did not reject the power-law hypothesis. Likelihood ratio tests favored the power law over the exponential ($R = 191.085$, with $p < 0.001$) and found no significant difference from the lognormal probability distribution ($R = -0.040$, with $p = 0.859$). The exponent lies in the critical regime $2 < \alpha < 3$, implying a finite mean but divergent theoretical variance, reflected in extreme sensitivity to tail values and unstable variance estimates. These features undermine conventional machine learning, leading to: (i) violations in methods assuming stable variance; (ii) under-representation of high-impact tail events in random samples; and (iii) unreliable extrapolation by synthetic oversampling methods. Our ensemble approach mitigates these issues by preserving the authentic minority class and ensuring tail representation.

Bootstrap analysis with $1,000$ iterations confirmed extreme loss concentration: 22.5% of loss-generating customers account for 80% of total losses, with a 95% confidence interval (CI) of $(20.7\%, 24.4\%)$. As shown in Table 1 and Figure 2, a minority of customers drive most losses, reinforcing the need to prioritize minority-class identification over global classification accuracy.

Table 1. Loss concentration across multiple thresholds.

Cumulative Loss	Customer Percentage	95% CI
70%	13.5%	(11.8%, 15.1%)
75%	17.4%	(15.6%, 19.1%)
80%	22.5%	(20.7%, 24.4%)
85%	29.4%	(27.5%, 31.2%)
90%	38.5%	(36.7%, 40.2%)

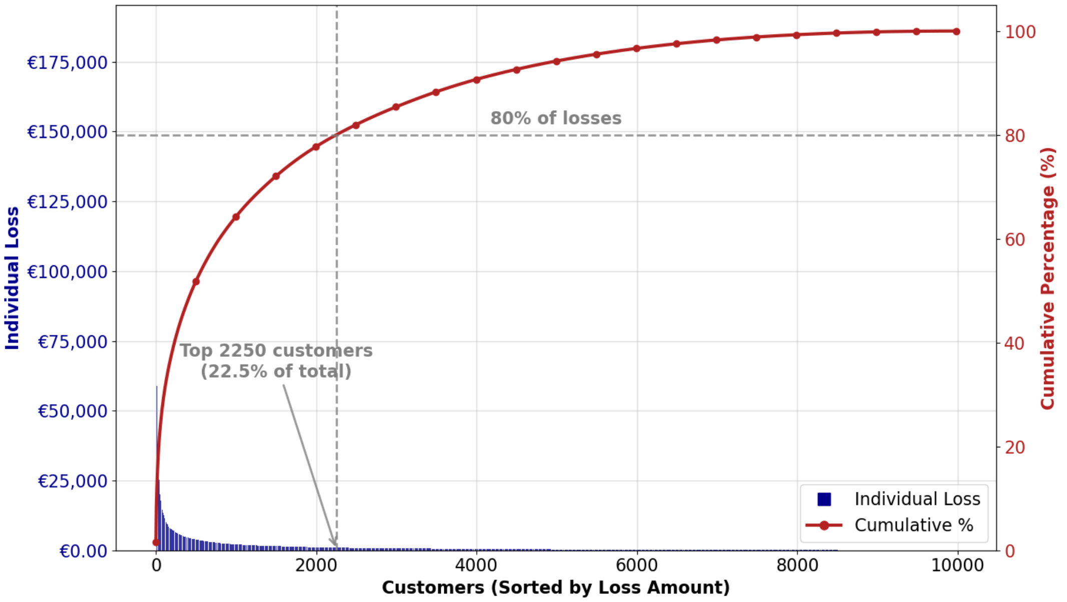

Figure 2. Cumulative loss distribution of customers.

4. Binary Classification Under Business Constraints

The binary classification process operates under strict business constraints reflecting real-world deployment conditions. The insurance company can omit at most 8% of potential customers due to competitive positioning, regulatory requirements, and acquisition targets. These customers are not rejected outright; instead, they undergo a deeper, personalized risk assessment by company experts. This limit is non-negotiable and defines the conditions under which any model must function. Rather than tuning this threshold to maximize validation metrics, which risks overfitting to our dataset, we adopt the business-imposed constraint directly. The classification threshold is set at the $8th$ percentile of the distribution of the target variable (customer's average annual benefit), corresponding to values below $-1,708$ euros. This choice emphasizes generalizability and practical applicability over dataset-specific performance gains.

This constraint produces a severe class imbalance with direct practical implications. The model flags $4,129$ customers (8% of the dataset) as high-loss cases, yielding a minority-to-majority ratio of about 1 to 11.5. Although the minority class represents only 8% of customers, it accounts for a disproportionate share of negative business impact. Standard machine learning methods trained on such imbalanced data often show strong overall accuracy while failing to identify the minority class, precisely the opposite of business requirements. The asymmetry of misclassification costs drives our methodological design. False negatives (missed high-loss customers) lead to direct financial losses, as these customers generate costs far exceeding their premiums. False positives (flagging profitable customers as high-risk) create opportunity costs but not direct losses. Our analysis shows that the $4,129$ high-loss customers in the dataset would produce losses far exceeding the forgone profits from omitting profitable customers. This supports the business logic of prioritizing minority class recall over overall accuracy.

5. Balanced Bagging Ensemble Approach

Our approach builds upon the established foundation of ensemble methods for class imbalance learning [44], while introducing novel elements specifically designed for heavy-tailed financial risk assessment. While methods such as EasyEnsemble [45] and SMOTE-Boost [46] have demonstrated effectiveness in general imbalanced learning scenarios, the specific requirements of financial risk assessment, including extreme loss concentration, business constraints, and asymmetric costs, motivate our specialized approach. Rather than attempting to compare against all existing methods, we focus on demonstrating the effectiveness of our business-constrained, profit-optimized ensemble approach through rigorous empirical validation against the results achieved in Soriano-Gonzalez et al. [7], which were implemented under identical conditions for the same dataset.

The theoretical motivation for balanced bagging in heavy-tailed loss scenarios stems from three key principles. Firstly, in power-law distributed losses ($\alpha = 2.381$), accurate identification of minority class instances yields disproportionate business value due to extreme loss concentration. Our analysis demonstrates that 22.6% of loss-generating customers account for 80% of total losses, creating a scenario where minority class precision directly translates to major risk mitigation. Secondly, unlike synthetic oversampling methods, our approach maintains the genuine characteristics of minority class instances. This is crucial for heavy-tailed distributions where the extreme tail behavior (which drives business risk) cannot be reliably synthesized without introducing distributional artifacts that may compromise risk assessment accuracy. Finally, by training each base learner on different random samples of the majority class while maintaining complete minority class coverage, we ensure in depth exploration of the majority class decision space while providing consistent minority class reinforcement.

Algorithm 1 shows the pseudo-code of our balanced-bagging algorithm. It constructs K base classifiers, where each classifier h_k is trained on a balanced subset $\mathcal{D}_k$ created through controlled sampling. It employs XGBoost [47] as the base learner due to its demonstrated effectiveness in handling non-linear relationships, feature interactions, and gradient-based optimization. XGBoost's tree-based architecture is particularly suitable for financial risk modeling as it naturally captures the hierarchical decision structures present in customer risk assessment.

The optimal sampling ratio r was determined through a systematic two-stage optimization process designed to balance computational efficiency with exploration. The

Algorithm 1 Balanced Bagging for Heavy-Tailed Loss Distribution

```
Require: Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, ensemble size $K$, sampling ratio $r$
Ensure: Ensemble $\mathcal{H} = \{h_1, h_2, \dots, h_K\}$
Partition $\mathcal{D}$ into $\mathcal{D}_{\text{maj}}$ and $\mathcal{D}_{\text{min}}$ based on class labels
for $k = 1$ to $K$ do
    Sample $\lfloor |\mathcal{D}_{\text{min}}| \cdot r \rfloor$ instances from $\mathcal{D}_{\text{maj}}$ to form $\mathcal{D}_{\text{maj}}^{(k)}$
    Create balanced subset $\mathcal{D}_k = \mathcal{D}_{\text{min}} \cup \mathcal{D}_{\text{maj}}^{(k)}$
    Train base classifier $h_k$ on $\mathcal{D}_k$ using XGBoost with fixed hyperparameters
    Add $h_k$ to ensemble $\mathcal{H}$
end for
return $\mathcal{H}$
```

first stage constitutes of conducting a preliminary random search across a broad range $r \in [1.0, 20.0]$ to identify promising regions. This exploratory phase revealed that performance gains decrease significantly beyond $r = 4$, while computational costs increase linearly with sampling ratio. The second stage is based on the exploratory results, we performed systematic grid search across the refined range ($r \in [1.0, 4.0]$), testing a discrete set of values for r: $r \in \{1.17, 1.47, 2.12, 2.80, 3.12, 3.20, 3.60, 3.85\}$. Each configuration was evaluated using hold-out validation on a stratified sample to ensure representative class distribution. The optimization framework employs a constrained profit maximization approach:

$$r^* = \arg\max_r \mathbb{E}[\text{Profit}_{\text{validation}}(r)] \tag{1}$$

$$\text{subject to} \quad \frac{|\{i : \hat{y}_i(r) = 1\}|}{n_{\text{validation}}} \leq 0.08 \tag{2}$$

The optimal sampling ratio was determined through 5-fold stratified cross-validation repeated 3 times ($n = 15$ evaluations per strategy) to ensure robust parameter selection. Confidence intervals were computed using non-parametric (bootstrap) methods. Statistical significance was assessed using Friedman tests for multiple comparisons ($\chi^2 = 98.67$, with $p < 0.0001$) and Wilcoxon signed-rank tests for pairwise comparisons ($\alpha = 0.05$). The profit function calculates realized benefit by omitting customers classified as high-loss (class 1) and summing the values of average annual benefit for selected customers (class 0):

$$\text{Profit}(r) = \sum_{i \in \mathcal{I}_{selected}(r)} \text{avg. annual profit}_i$$

where $\mathcal{I}_{selected}(r) = \{i : \hat{y}_i(r) = 0\}$ represents the set of customers selected under sampling ratio r. This profit-centric objective directly aligns with business value creation while the constraint ensures adherence to operational limitations imposed by competitive market positioning and regulatory compliance requirements.

Final predictions are generated through majority voting among the K base classifiers:

$$\hat{y} = \arg\max_{c \in \{0,1\}} \sum_{k=1}^{K} \mathbf{1}[h_k(x) = c]$$

where $\mathbf{1}[\cdot]$ denotes the indicator function. This aggregation strategy provides natural uncertainty quantification through vote distribution and enhances robustness against individual classifier errors, which becomes a critical consideration when model decisions directly impact financial outcomes. The ensemble size $K = 50$ was selected based on convergence analysis showing that additional base learners beyond this threshold provide marginal performance improvements while increasing computational overhead. This configuration balances prediction stability with computational efficiency for practical deployment scenarios.

6. Sampling Strategy Optimization Results

The random search optimization evaluated eight different sampling strategies across the range $r \in [1.17, 3.85]$, with each configuration assessed through 5-fold stratified cross-validation repeated 3 times ($n = 15$ evaluations per strategy). Bootstrap resampling with $1,000$ iterations provided robust confidence interval estimation for all performance metrics. Table 2 presents a performance analysis for each sampling ratio.

Table 2. Random search results for sampling strategy optimization.

Sampling Strategy	Mean Real Benefit (€)	Bootstrap 95% CI (€)	Clients Omitted (%)	F1-Score (mean)	Constraint Compliant
1.17	727,833	$(687621, 766880)$	27.6	0.797	No
1.47	713,699	$(674201, 754649)$	20.5	0.836	No
2.12	616,093	$(561111, 671588)$	12.4	0.874	No
2.80	536,471	$(472676, 595574)$	8.2	0.890	No
3.12	505,111	$(436844, 566107)$	7.0	0.894	Yes
3.20	489,752	$(422913, 553440)$	6.8	0.894	Yes
3.60	460,843	$(399643, 526616)$	5.7	0.897	Yes
3.85	435,281	$(382279, 491054)$	5.1	0.898	Yes

Statistical significance testing revealed large differences between strategies. The Friedman test confirmed significant variation across all strategies ($\chi^2 = 98.67$, with $p < 0.0001$). While sampling ratios 1.17 and 1.47 achieved the highest absolute profits ($727,833$ euros and $713,699$ euros, respectively), these strategies violated the operational constraint by omitting 28% and 21% of customers, respectively. Among constraint-compliant strategies ($\leq 8\%$ omission), sampling ratio $r^* = 3.12$ emerged as a good candidate, achieving the highest real benefit ($505,111$ euros, with bootstrap 95% confidence interval of $(436844, 566107)$), while maintaining 7.0% customer omission and an F1-score of 0.894. Although sampling ratio 2.80 achieved higher absolute profit ($536,471$ euros), it violated the operational constraint with 8.2% customer omission, rendering it unsuitable for deployment.

7. Model Performance on Validation and Test Sets

The optimized balanced ensemble with $r = 3.12$ demonstrated robust performance on the validation set, which was created from the training dataset and comprised $7,227$ customers. Figure 3 summarizes the classification performance metrics.

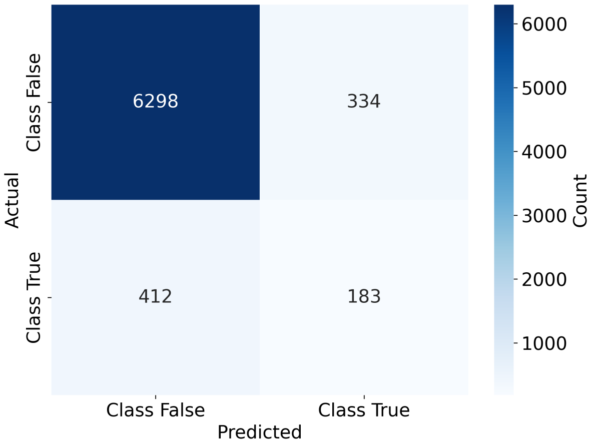

Figure 3. Confusion matrix of the balanced bagging model for the validation set.

The confusion matrix revealed $6,298$ true negatives, 183 true positives, 334 false positives, and 412 false negatives. While the model achieved excellent performance on the majority class (low-risk customers), minority class detection presented expected challenges due to the severe class imbalance and limited historical data availability. From a business perspective, the model created value. The validation set achieved a real benefit of $701,436$ euros, representing 41% of the theoretical maximum benefit ($1,728,523$ euros). This translates to an average benefit of 104.54 euros per customer while omitting exactly 7.15% of the customer base, well within the operational constraint. The final model evaluation on the independent test set of $15,486$ customers confirmed the robustness and generalizability of the balanced ensemble approach. Figure 4 presents the comprehensive test set performance metrics.

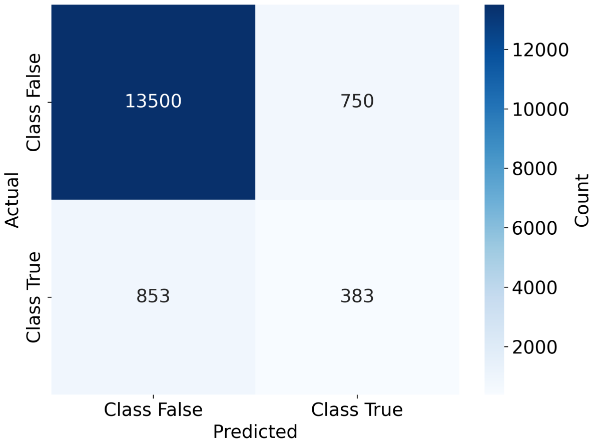

Figure 4. Confusion matrix of the balanced bagging model for the test set.

The test set confusion matrix revealed $13,500$ true negatives, 383 true positives, 750 false positives, and 853 false negatives. Performance metrics remained consistent with validation results, indicating stable model behavior across different data partitions. The business impact analysis reveals important value creation. The model achieved a real benefit of $1,329,076$ euros on the test set, representing 36% of the theoretical maximum benefit ($3,674,846$ euros). This corresponds to an average benefit of 92.60 euros per customer while omitting 7.32% of potential customers, maintaining strict adherence to operational constraints.

8. Economic Impact and Misclassification Analysis

The detailed analysis of test set predictions reveals economic trade-offs that are inherent in customer classification under business constraints. Table 3 presents the comprehensive breakdown of misclassification costs and their business implications.

Table 3. Economic impact analysis by customer category.

Category	Count	Mean Profit (€)	Total Impact (€)	Business Implication
True Negatives	13,500	244.89	$+3,306,019$	Correctly Selected
True Positives	383	$-2,606.71$	$-998,371$	Correctly Omitted
False Negatives	853	$-2,317.64$	$-1,976,944$	Missed high-risk
False Positives	750	100.40	$+75,302$	Foregone profit

The asymmetric cost structure validates the model's conservative approach. False negatives impose much higher per-customer costs ($-2,318$ euros average loss) compared to false positives (100 euros average foregone profit), justifying the emphasis on precision over recall in minority class detection. Analysis of the most severe misclassifications reveals concentrated risk exposure. The ten worst false negatives account for 24.4% of total false negative losses ($-481,889$ euros out of $-1,976,944$ euros), with the single worst case representing a $-176,973$ euros loss. This concentration effect demonstrates that while the model successfully identifies the majority of high-risk customers, the most extreme cases remain challenging to detect with available features.

Conversely, false positive analysis shows more distributed impact. The ten highest-value omitted customers represent 23.3% of foregone profit ($17,534$ euros out of $75,302$ euros), with the maximum individual opportunity cost of $4,393$ euros. The relatively modest false positive impact, combined with their distributed nature, suggests that omission errors impose manageable opportunity costs compared to selection errors. The model demonstrates strong performance in risk concentration mitigation. True positives exhibit an average loss of $-2,607$ euros per customer, while correctly selected customers (true negatives) show an average profit of 245 euros per customer. This $2,852$ euros average differential per correctly classified customer underscores the value creation potential of accurate minority class identification. The sensitivity analysis reveals that false negatives concentrate in the $-1,000$ to $-2,000$ loss range ($25th$ to $75th$ percentiles), suggesting systematic challenges in distinguishing moderate-loss customers from profitable ones. However, the model successfully captures extreme tail risk, as evidenced by the omission of 383 high-risk customers with total expected losses of $998,371$ euros.

9. Comparative Analysis with Baseline Methodology using a Real Dataset

To evaluate the effectiveness of our balanced ensemble approach, we compare our results with the methodology proposed by Soriano-Gonzalez et al. [7], who applied traditional gradient boosting methods (XGBoost and LightGBM) with early stopping to the same customer classification problem. Table 4 presents a comprehensive comparison between our balanced ensemble approach and the aforementioned baseline methodology.

Table 4. Performance comparison: balanced ensemble vs. baseline methodology.

Metric	Soriano-Gonzalez et al. [7]	Our Approach	Improvement
ROC-AUC (Test)	0.72	0.90	+25.0%
Precision (High-Risk)	0.67	0.34	−49.3%
Recall (High-Risk)	0.23	0.31	+34.8%
F1-Score (Weighted)	0.79	0.89	+12.7%
Customer Omission Rate	6.0%	7.0%	$+1.0pp$
Business Metrics			
Test Set Benefit	1,232,663 €	1,329,076 €	+7.8%
Avg. Benefit per Customer	85 €	93 €	+9.4%
Benefit as % of Maximum	36%	36%	0pp

The comparison reveals several insights about the effectiveness of balanced ensemble methods versus traditional approaches. While Soriano-Gonzalez et al. applied traditional gradient boosting methods with early stopping and used a simple binary threshold based on positive versus negative average annual profit, our balanced ensemble approach explicitly addresses class imbalance through systematic undersampling with ensemble aggregation. This fundamental difference in approach results in improved overall accuracy (90% vs 83%) and weighted F1-score (0.89 vs 0.79). The methodologies exhibit contrasting performance characteristics in minority class detection. The baseline methodology achieved higher precision for high-risk customers (67% vs 34%), indicating fewer false positive errors. However, our approach demonstrates superior recall (31% vs 23%), successfully identifying more actual high-risk customers. This trade-off favors our approach given the asymmetric cost structure, where false negatives impose higher costs ($-2,318$ euros average) than false positives (100 euros average). Both methodologies respect operational constraints, with omission rates of 6.0% (baseline) and 7.0% (our approach), both well within the 8% user-defined threshold. However, our approach achieves $96,413$ additional profit while maintaining constraint compliance. The superior recall of our method translates to identification of 68 additional high-risk customers (31% vs 23% of $1,236$ actual high-risk customers), potentially avoiding 157, 624 euros in losses based on the average high-risk customer loss of −2, 318 euros.

While our balanced ensemble methodology demonstrates improvements over baseline approaches, several limitations warrant critical examination. The most significant limitation is the large reduction in precision for high-risk customer identification (34% vs 67% baseline). This results in 750 false positives compared to an estimated 326 in the baseline approach, representing an additional 424 profitable customers incorrectly omitted. At 100 eruos average foregone profit per false positive, this translates to 42, 400 in additional opportunity costs that partially offset the method's benefits. The balanced ensemble approach systematically discards majority class information, using only 33% of available low-risk customer data in each base learner (sampling ratio $r = 3.12$). This information loss may explain the reduced precision, as the model receives limited exposure to the full diversity of profitable customer patterns. Alternative approaches such as cost-sensitive learning or synthetic oversampling might preserve this information while maintaining class balance.

From a computational perspective, training 50 ensemble models, each on balanced subsets, requires significantly more computational resources than single-model approaches. The ensemble's computational cost represents a relatively large increase for relatively modest performance gains (7.8% profit improvement), raising questions about scalability for production deployment. Our approach used fixed XGBoost hyper-parameters across all ensemble members, potentially leaving performance gains unrealized. The baseline methodology employed hyper-parameter optimization, which may partially explain their superior precision. While we optimized sampling ratios, we used majority voting for final predictions rather than optimizing decision thresholds. The baseline approach's threshold optimization at 0.5 may be suboptimal for this cost structure, but our method lacks explicit threshold calibration, potentially limiting performance ceiling. The optimal sampling ratio ($r = 3.12$) was derived from our specific dataset's characteristics.

10. Comparative Analysis with Baseline Methodology using a Synthetic Dataset

To enable reproducibility and support future methodological comparisons, we repeat the evaluation using a synthetic dataset derived from the original real data. This dataset has been generated by introducing controlled random noise into the float variables of the original dataset. The noise was normally distributed and added in such a way that the core structure and statistical patterns of the original variables were preserved. To ensure realism, synthetic values were clipped to remain within the original variable ranges. This process maintains the general relationships between variables while protecting the original data.

We apply the same baseline methodology described in Soriano-Gonzalez et al. [7], including the use of XGBoost and LightGBM with early stopping, to this synthetic dataset. Our balanced ensemble approach is then applied under the same experimental setup. The objective is to assess whether the relative performance gains observed in Section 9 are consistent when applied to public, shareable data. Table 5 presents the comprehensive performance comparison on the synthetic test dataset containing 15, 486 customers.

Table 5. Performance comparison on synthetic dataset: balanced ensemble vs. baseline methodology.

Metric	Soriano-Gonzalez et al. [7]	Our Approach	Improvement
ROC-AUC (Test)	0.70	0.76	+8.1%
Precision (High-Risk)	67%	35%	−47.8%
Recall (High-Risk)	19%	28%	+47.4%
F1-Score (Weighted)	0.79	0.90	+13.9%
Customer Omission Rate	6.0%	6.35%	+0.35pp
Business Metrics			
Test Set Benefit	1,074,418 €	1,245,861 €	+16.0%
Avg. Benefit per Customer	69 €	86 €	+23.8%
Benefit as % of Maximum	29%	34%	$+5pp$

The synthetic dataset results confirm the robustness of our balanced ensemble approach and reveal similar performance trade-offs as observed with the real dataset. Our method achieves improvements in overall accuracy (90% vs. 83%) and ROC-AUC (0.76 vs. 0.70). The weighted F1-score improvement of 13.9% (0.90 vs. 0.79) indicates better overall classification performance across both classes.

For minority class detection, the results mirror the trade-offs observed in Section 9. While the baseline methodology achieves higher precision for high-risk customers (67% vs. 35%), our balanced ensemble approach demonstrates superior recall (28% vs. 19%), successfully identifying 47.4% more actual high-risk customers. This translates to detecting approximately 111 additional high-risk customers of $1,236$ total high-risk customers, which is crucial given the asymmetric cost structure where false negatives impose significantly higher costs than false positives.

The business impact analysis reveals value creation. Our approach achieves a 16% improvement in total benefit ($1,245,861$ euros vs. $1,074,418$ euros) and a 23.8% increase in average benefit per customer (85.91 euros vs. 69.38 euros). Both approaches maintain similar omission rates within operational constraints, with our method excluding 6.4% of customers versus 6.0% for the baseline. Notice that the performance gains on the synthetic dataset are more pronounced than those observed on the real dataset (16.0% vs. 7.8% improvement in total benefit), which may result from the noise introduction process amplifying class separation characteristics that favor ensemble-based approaches.

The consistency of the precision-recall trade-off across both real and synthetic datasets validates the fundamental characteristics of our approach: improved overall accuracy and recall at the cost of reduced precision in minority class detection. This trade-off proves economically favorable given the asymmetric cost structure inherent in insurance risk assessment. The synthetic dataset results provide a reproducible benchmark for future methodological comparisons in insurance risk classification, enabling researchers to evaluate alternative approaches against our established baseline while confirming the practical applicability of balanced ensemble methods across diverse data environments.

11. Conclusion

This study presents a classification method for new insurance customers without historical data, combining balanced bagging ensembles with a profit-oriented objective function under explicit operational constraints. The approach addresses severe class imbalance and heavy-tailed loss distributions by preserving all minority-class cases and systematically undersampling the majority class. The model integrates authentic data distribution, profit-based thresholding, and ensemble majority voting to balance business value creation with deployment feasibility.

Applied to a real auto insurance dataset of $51,618$ customers, the method identified 31% of actual high-risk customers, improving recall over the baseline by 34.8% while keeping customer omission within the fixed 8% limit. The optimal sampling ratio ($r = 3.12$) produced a real test-set benefit of $1,329,076$ euros, equal to 36% of the theoretical maximum, and an average profit of 92.60 euros per customer. This represented a $96,413$ euros improvement over the benchmark while omitting 7.3% of customers. Gains were consistent on a synthetic dataset, where the model increased total benefit by 16% and average profit per customer by 23.8% compared to the baseline.

The approach reduced exposure to concentrated losses, with true positives averaging $-2,607$ euros in avoided losses per customer, and demonstrated robustness across validation and independent test sets. The design choices (undersampling to maintain minority coverage, using multiple base learners for decision stability, and profit-maximizing under constraint) produced a model that aligns predictive performance with economic priorities. The precision–recall trade-off, favoring higher recall of high-loss cases at the cost of more false positives, proved economically justified given the asymmetric cost structure, where false negatives were on average more than 20 times more costly than false positives. The methodology can be deployed without violating business acceptance constraints and is adaptable to other insurance lines with similar imbalance and cost patterns.

Future work will focus on reducing computational demands of the ensemble while maintaining performance, exploring cost-sensitive learning to recover some precision, and extending validation to other markets and regulatory environments. Integrating adaptive mechanisms for threshold adjustment in response to market or portfolio changes may further strengthen long-term applicability.

Author Contributions: Conceptualization A.A.J.; methodology, F.L.S.; software, F.L.S. and A.A.J.; validation, R.S.; formal analysis, R.S.; data curation, F.L.S.; writing—original draft preparation, F.L.S. and R.S.; writing—review and editing, A.A.J.; supervision, R.S. and A.A.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been partially funded by the Spanish Ministry of Science and Innovation (PID2022-138860NB-I00, RED2022-134703-T).

Data Availability Statement: The benchmark dataset presented in this paper is available at https://www.researchgate.net/publication/394469850_car_insurance_initial_customer_segmentation_dataset.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AUC	Area Under the Curve
CI	Confidence Interval
ML	Machine Learning
MSE	Mean Squared Error
ROC AUC	Receiver Operating Characteristic – Area Under the Curve
SINCO	Information System of the Insurance Compensation Consortium in Spain
SMOTE	Synthetic Minority Over-sampling Technique

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.

References

1. Finger, D.; Albrecher, H.; Wilhelmy, L. On the cost of risk misspecification in insurance pricing. *Japanese Journal of Statistics and Data Science* **2024**, *7*, 1111–1153.
2. Bruun, S.B.; Lioma, C.; Maistro, M. Recommending Target Actions Outside Sessions in the Data-poor Insurance Domain. *ACM Transactions on Recommender Systems* **2024**, *3*, 1–24.
3. Hu, C.; Quan, Z.; Chong, W.F. Imbalanced learning for insurance using modified loss functions in tree-based models. *Insurance: Mathematics and Economics* **2022**, *106*, 13–32.
4. Singla, J.; Bashir, A.K.; Nam, Y.; Hasan, N.U.; Tariq, U.; et al. Handling class imbalance in online transaction fraud detection. *Computers, Materials and Continua* **2021**, *70*, 2861–2877.
5. Ngwenduna, F.; et al.. Alleviating Class Imbalance in Actuarial Applications Using Cost-Sensitive Learning. *Risks* **2021**, *9*, 49.
6. Martin, J.; Taheri, S.; Abdollahian, M. Optimizing ensemble learning to reduce misclassification costs in credit risk scorecards. *Mathematics* **2024**, *12*, 855.
7. Soriano-Gonzalez, R.; Tsertsvadze, V.; Osorio, C.; Fuster, N.; Juan, A.A.; Perez-Bernabeu, E. Balancing Risk and Profit: Predicting the Performance of Potential New Customers in the Insurance Industry. *Information* **2024**, *15*, 546.
8. Uddin, M.; et al.. Modeling Vehicle Insurance Adoption by Automobile Owners Using PCA and SMOTE. *Processes* **2023**, *11*, 90.
9. Marhavilas, P.K.; Koulouriotis, D.E. Risk-acceptance criteria in occupational health and safety risk-assessment—the state-of-the-art through a systematic literature review. *Safety* **2021**, *7*, 77.
10. Leo, M.; Sharma, S.; Maddulety, K. Machine learning in banking risk management: A literature review. *Risks* **2019**, *7*, 29.

11. Fitriani, M.A.; Febrianto, D.C. Data mining for potential customer segmentation in the marketing bank dataset. *JUITA: Jurnal Informatika* **2021**, *9*, 25–32.
12. Simester, D.; Timoshenko, A.; Zoumpoulis, S.I. Targeting prospective customers: Robustness of machine-learning methods to typical data challenges. *Management Science* **2020**, *66*, 2495–2522.
13. Hutagaol, B.J.; Mauritsius, T. Risk level prediction of life insurance applicant using machine learning. *International Journal of Advanced Trends in Computer Science and Engineering* **2020**, *9*.
14. Sahai, R.; Al-Ataby, A.; Assi, S.; Jayabalan, M.; Liatsis, P.; Loy, C.K.; Al-Hamid, A.; Al-Sudani, S.; Alamran, M.; Kolivand, H. Insurance risk prediction using machine learning. In Proceedings of the The international conference on data science and emerging technologies. Springer, 2022, pp. 419–433.
15. Henckaerts, R.; Côté, M.P.; Antonio, K.; Verbelen, R. Boosting insights in insurance tariff plans with tree-based machine learning methods. *North American Actuarial Journal* **2021**, *25*, 255–285.
16. Dietterich, T.G. Ensemble methods in machine learning. In Proceedings of the International workshop on multiple classifier systems. Springer, 2000, pp. 1–15.
17. Zhou, Z.H. *Ensemble methods: foundations and algorithms*; CRC press, 2012.
18. Brati, E.; Braimllari, A.; Gjeçi, A. Machine Learning Applications for Predicting High-Cost Claims Using Insurance Data. *Data* **2025**, *10*, 90.
19. Krupova, M.; Rachdi, N.; Guibert, Q. Explainable boosting machine for predicting claim severity and frequency in car insurance. *arXiv preprint arXiv:2503.21321* **2025**.
20. Hanafy, H.; et al.. Machine Learning Approaches for Auto Insurance Big Data. *Risks* **2021**, *9*, 106.
21. Baran, S.; Rola, P. Prediction of motor insurance claims occurrence as an imbalanced machine learning problem. *arXiv preprint arXiv:2204.06109* **2022**.
22. Khamesian, F.; Esna-Ashari, M.; Dei Ofosu-Hene, E.; Khanizadeh, F. Risk Classification of Imbalanced Data for Car Insurance Companies: Machine Learning Approaches. *International Journal of Mathematical Modelling & Computations* **2022**, *12*, 153–162.
23. Gupta, V.; Broughton, J.; Rukundo, A.; Pinky, L. Learning Unbiased Risk Prediction Based Algorithms in Healthcare: A Case Study with Primary Care Patients. *Available at SSRN 4984535*.
24. Li, Y.; Zhang, Y.; Gan, L.; Hong, G.; Zhou, Z.; Li, Q. RevMan: Revenue-aware multi-task online insurance recommendation. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2021, Vol. 35, pp. 303–310.
25. Loisel, S.; Piette, P.; Tsai, C.H.J. Applying economic measures to lapse risk management with machine learning approaches. *ASTIN Bulletin: The Journal of the IAA* **2021**, *51*, 839–871.
26. Sadreddini, Z.; Donmez, I.; Yanikomeroglu, H. Cancel-for-Any-Reason Insurance Recommendation Using Customer Transaction-Based Clustering. *IEEE Access* **2021**, *9*, 39363–39374.
27. Sari, P.K.; Purwadinata, A. Analysis characteristics of car sales in E-commerce data using clustering model. *Journal of Data Science and Its Applications* **2019**, *2*, 19–28.
28. Tian, X.; Todorovic, J.; Todorovic, Z. A Machine-Learning-Based Business Analytical System for Insurance Customer Relationship Management and Cross-Selling. *Journal of Applied Business & Economics* **2023**, *25*.
29. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *nature* **2015**, *521*, 436–444.
30. Elbhrawy, A.S.; Belal, M.A.; Hassanein, M.S. CES: Cost Estimation System for Enhancing the Processing of Car Insurance Claims. *Journal of Computing and Communication* **2024**, *3*, 55–69.
31. De Meulemeester, H.; De Moor, B. Unsupervised embeddings for categorical variables. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020, pp. 1–8.
32. Kolambe, S.; Kaur, P. Survey on Insurance Claim analysis using Natural Language Processing and Machine Learning. *International Journal on Recent and Innovation Trends in Computing and Communication* **2023**, *11*, 30–38.
33. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine* **2018**, *13*, 55–75.
34. Cambria, E.; White, B. Jumping NLP curves: A review of natural language processing research. *IEEE Computational intelligence magazine* **2014**, *9*, 48–57.
35. Orji, U.; Ukwandu, E. Machine learning for an explainable cost prediction of medical insurance. *Machine Learning with Applications* **2024**, *15*, 100516.
36. Doshi-Velez, F.; Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* **2017**.
37. Le, T.T.H.; Prihatno, A.T.; Oktian, Y.E.; Kang, H.; Kim, H. Exploring local explanation of practical industrial AI applications: A systematic literature review. *Applied Sciences* **2023**, *13*, 5809.

38. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Advances in neural information processing systems* **2017**, *30*.
39. Sharma, A.; et al. Demystifying Privacy-preserving AI: Strategies for Responsible Data Handling. *MZ Journal of Artificial Intelligence* **2024**, *1*, 1–8.
40. Voigt, P.; Von dem Bussche, A. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing* **2017**, *10*, 10–5555.
41. Jarque, C.M.; Bera, A.K. A test for normality of observations and regression residuals. *International Statistical Review* **1987**, *55*, 163–172.
42. Anderson, T.W.; Darling, D.A. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics* **1952**, *23*, 193–212.
43. Clauset, A.; Shalizi, C.R.; Newman, M.E. Power-law distributions in empirical data. *SIAM review* **2009**, *51*, 661–703.
44. Galar, M.; Fernandez, A.; Barrenechea, E.; Bustince, H.; Herrera, F. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* **2012**, *42*, 463–484.
45. Liu, X.Y.; Wu, J.; Zhou, Z.H. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **2009**, *39*, 539–550.
46. Chawla, N.V.; Lazarevic, A.; Hall, L.O.; Bowyer, K.W. SMOTEBoost: Improving prediction of the minority class in boosting. In Proceedings of the European conference on principles of data mining and knowledge discovery. Springer, 2003, pp. 107–119.
47. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.