Business-Constrained Risk Prediction for Insurance: A macine learning approach for customer classification

Finn Luca Solly[1][0009−0008−2993−4284], Raquel Soriano-Gonzalez[2][0000−0002−1337−9561], and Angel A. Juan[2,3][0000−0003−1392−1776]

[1] ESADE Business School, Barcelona, Spain finnluca.solly@alumni.esade.edu
[2] CIGIP, Universitat Politècnica de València, Alcoy, Spain
rsorgon@epsa.upv.es,ajuanp@upv.es
[3] Euncet Business School, Terrassa, Spain
ajuanp@upv.es

Abstract. Customer classification in the insurance industry poses unique challenges due to extreme class imbalance, heavy-tailed loss distributions, and strict operational constraints. Traditional machine learning models often fail to identify high-risk clients effectively, as they prioritize global accuracy over business-critical outcomes. In this study, we propose a business-aware classification framework based on a balanced bagging ensemble, specifically designed to address the asymmetric cost structure and exclusion limits present in real-world insurance settings. Our method combines authentic preservation of the minority class with systematic optimization of the majority sampling ratio, using a profit-based objective function constrained by a maximum customer exclusion threshold. Applied to a large-scale auto insurance dataset, the proposed approach significantly improves the detection of loss-generating customers and maximizes net business value while ensuring regulatory and market compliance. This work contributes to the growing field of business-optimized machine learning and offers a practical and robust solution for risk-driven decision-making in the insurance sector.

Keywords: Insurance Risk Modeling · High-Risk Customer Detection · Profit-Aware Machine Learning.

1 Introduction

In the insurance industry, accurate customer risk assessment is fundamental to maintaining profitability and competitive advantage. Studies indicate that misclassification of customers can result in considerable financial losses, with poor risk assessment contributing to a significant increase in losses that impact business performance [1]. The challenge of predicting customer performance becomes particularly difficult when dealing with potential new customers, where available historical data is limited and the cost of misclassifications can be substantial [2].

However, customer classification in insurance presents a technical challenge due to the severe class imbalance, where loss-generating customers traditionally only represent 10–20% of the customer base but can still account for 60–80% of total losses [3]. Traditional binary classification approaches struggle with such highly imbalanced datasets, often resulting in models that are biased towards the majority class and fail to correctly identify high-loss customers [4]. This class imbalance problem is augmented by practical business constraints, as insurance companies face limitations on the percentage of customers that can be excluded without jeopardizing business volume or market share.

The financial consequences of misclassification have a critical impact, specifically in insurance contexts. False negatives, i.e., misclassifying high-risk customers as low-risk, result in direct financial losses through underpriced premiums that do not cover future claims, whereas false positives (rejecting profitable customers) lead to opportunity costs [5]. This trade-off calls for sophisticated approaches that can effectively balance risk identification with business sustainability requirements.

When addressing these customer classification issues, the existing literature has mainly focused on the application of feature selection and model optimization. Soriano-Gonzalez et al. [6] investigated the prediction of potential new customer performance using models such as XGBoost and LightGBM, supported by extensive feature engineering and profitability-based thresholds. However, their approach does not explicitly address the severe class imbalance or the asymmetric costs associated with different types of misclassification. Furthermore, Soriano-Gonzalez et al. rely on arbitrary economic thresholds and do not incorporate operational constraints such as acceptable rejection rates or business risk tolerance. Moreover, their study does not explore ensemble strategies specifically designed to mitigate class imbalance, nor does it consider business-driven principles, such as the Pareto rule, to guide model thresholds.

Against this background, the primary goal of this paper is to develop a classification framework that integrates a balanced ensemble approach with business-aware thresholding, explicitly designed to manage class imbalance and maintain practical exclusion constraints. Our approach proposes to use the Pareto principle to identify the optimal threshold for high-loss customer classification, coupled with a balanced ensemble of XGBoost models trained on randomly sampled subsets of the data, combining business insight with advanced machine learning techniques. We use grid search to optimize the sampling and ensemble parameters. The methodology applied and results obtained in this paper are compared with the methodology proposed by Soriano-Gonzalez et al., using their approach as a performance benchmark to evaluate the effectiveness of balanced ensemble methods in contrast to traditional supervised learning classifiers.

In this study, we present a customer classification model for the insurance sector that combines ensemble learning techniques with business-oriented criteria. Throughout the paper, we describe the problem and review related work (Section 2), detail our methodology and model optimization process (Section 3), present the results obtained through various tests and datasets (Section 4), and

finally, summarize our conclusions and propose potential future improvements (Section 5).

2 Literature Review

Numerous recent studies have addressed risk prediction in the insurance sector using classical machine learning models. For example, research in life and auto insurance compares logistic regression, decision trees, Random Forest, and XGBoost, with XGBoost generally achieving superior performance (AUC 0.86) in customer risk classification tasks [7]. Another recent line of work leverages interpretable models such as Explainable Boosting Machines to predict claim frequency and severity, combining high accuracy with model transparency [8]. In addition, specialized models for high-cost auto claims have been developed to improve the early detection of high-impact cases [9].

Class imbalance is a common issue in insurance, where risk events, such as fraud or severe claims, are rare compared to the general population. Recent studies address this challenge by comparing techniques such as undersampling, oversampling (e.g., SMOTE), class weighting, and adaptive ensemble methods. In another study, Baran and Rola (2022) applied various algorithms, including decision trees and neural networks, in combination with resampling strategies to improve the detection of rare insurance claims in the auto sector [10]. Other research, particularly in the context of bias and fairness in medical risk models, has emphasized the importance of addressing class imbalance both ethically and technically in high-risk classification tasks [11].

Despite growing interest in machine learning applications within the insurance industry, few studies explicitly incorporate business constraints such as asymmetric misclassification costs or minimum acceptance thresholds. Although some recommendation-focused approaches—particularly on online insurance platforms—leverage multitask neural networks and reinforcement learning to optimize customer conversion through business-oriented logic [12], these are typically limited to digital distribution contexts. The broad adoption of machine learning for underwriting and pricing has been documented by organizations such as the OECD, yet this deployment often lacks direct integration of profit-driven objectives or operational constraints into model training pipelines [13].

In recent years, the use of advanced machine learning models for risk prediction in insurance has expanded significantly. However, many studies still exhibit critical limitations when it comes to systematically incorporating real-world business requirements and operational constraints into model design. A considerable number of recent works—such as those by Sahai et al. (2023) and Baran and Rola (2022)—employ powerful algorithms like XGBoost, Random Forest, or neural networks in insurance contexts with imbalanced data, yet they primarily optimize statistical metrics (AUC, precision, F1-score) without explicitly addressing asymmetric misclassification costs, exclusion constraints, or profit-driven objective functions. In addition, they often rely on fixed decision thresholds (e.g.,

$p = 0.5$) without economic or regulatory justification, which undermines the practical deployability of the models.

An important exception is the study by Soriano-González et al. (2024), which introduces an objective function focused on maximizing expected net profit and incorporates an economic analysis of misclassification costs in an insurance context. Nevertheless, this approach also presents several limitations: (i) although the presence of class imbalance is acknowledged, no specific technique is applied to mitigate it, which could hinder the model's ability to detect high-risk (minority) customers; (ii) a binary threshold based on expected gain or loss is used—while this aligns with financial reasoning, it does not explicitly incorporate business constraints such as a maximum allowable exclusion rate; (iii) there is no systematic optimization of the threshold or the structure of the training dataset to balance exclusion, profitability, and computational cost; and (iv) the approach relies on a single boosting model, without leveraging stratified subsampling ensemble techniques that could enhance robustness and allow for a broader exploration of the decision space.

Our study builds upon and refines the approach proposed by Soriano-González et al., introducing several key elements that provide additional value from both technical and business perspectives. We apply a Balanced Bagging strategy with multiple base classifiers, explicitly addressing the severe class imbalance present in the data while preserving the authentic structure of the minority class and avoiding synthetic oversampling artifacts. An explicit operational constraint—limiting customer exclusion to a maximum of 8%—is incorporated to ensure the commercial viability of the model and align it with real-world insurance industry practices. Furthermore, we conduct a systematic optimization of the sampling ratio (r) as a critical hyperparameter, employing stratified cross-validation and economic sensitivity analysis, which were not considered in previous work. Although we share with Soriano-González et al. the use of a profit-based objective function, our formulation explicitly defines the problem as a constrained maximization task, enabling direct control over the trade-off between profitability and adherence to business rules. Lastly, we add an additional layer of robustness through ensemble majority voting, enhancing both the stability of predictions and the model's ability to generalize across varying data conditions.

Overall, our approach complements and advances previous methodologies by offering a comprehensive solution that integrates powerful modeling techniques, economically meaningful metrics, and realistic business constraints, making it a practical tool for decision-making in real-world insurance settings.

3 Methodology

Our methodology consists of three main steps: data preparation and statistical analysis, defining class imbalance within business constraints, and developing a balanced ensemble classifier. This systematic approach addresses the fundamental challenges of predicting customer performance in scenarios characterized by extreme class imbalance and heavy-tailed loss distributions.

3.1 Data Preparation and Statistical Foundation

The empirical analysis utilizes a real-world insurance dataset comprising 116,934 customer records spanning the period 2016-2023. The initial dataset encompasses 196 features including demographic characteristics, vehicle attributes, policy details, and historical claims information. To ensure the methodology's applicability to prospective customers, we systematically removed features unavailable at the time of customer acquisition, including *PrimaTotalPoliza*, *ComisionTotalPoliza*, and related policy-specific variables.

Following Soriano-Gonzalez at al. [6], we applied standard preprocessing steps. Unicode normalization and column name standardization were implemented to ensure computational compatibility. We imputed missing values using SoftImputer, an iterative imputation method based on matrix completion. Customers lacking SINCO (Sistema de Información del Consorcio de Compensación de Seguros) data were excluded, resulting in a final dataset of 51,618 observations with 196 features.

To address multicollinearity concerns, we conducted correlation analysis using Pearson correlation coefficients. When features showed correlation > 0.8, we retained only the feature most correlated with the target variable (BMA_corregido) from each correlated group. This methodology resulted in the removal of 54 features from the original 196, yielding a final feature set of 142 variables while preserving the essential predictive information content of the dataset.

The target variable, BMA_corregido (corrected average annual margin), exhibits characteristics that fundamentally challenge standard machine learning assumptions. The Jarque-Bera test [14] statistic of 1.81×10^8 ($p < 0.001$) strongly rejects normality, while Anderson-Darling tests [15] reject all conventional distributions including normal, exponential, logistic, and Gumbel at all significance levels.

The loss distribution exhibits heavy tails, creating significant business risk concentration. Of the 51,618 customers in our dataset, 9,984 (19.3%) generate negative returns, representing a substantial minority that drives disproportionate business risk. These loss-generating customers account for total losses of €11,328,603, while the entire dataset generates a real benefit of only €909,881 before any risk management intervention. This stark contrast, where negative customers impose losses exceeding €11.3 million while the net portfolio benefit is less than €1 million, illustrates the critical importance of accurate loss customer identification. Following Clauset et al.[16], we conducted power-law analysis to characterize the extreme tail behavior driving business risk, employing maximum likelihood estimation resulting in a power-law exponent $\alpha = 2.381 \pm 0.057$ with lower threshold $x_{\min} = 3,454$. The Kolmogorov-Smirnov goodness-of-fit test produces a D-statistic of 0.0197 with $p = 0.652$, indicating the power-law model fits the data well, while the semi-parametric bootstrap approach cannot reject the power-law hypothesis. Model comparison tests support the power-law hypothesis over simpler alternatives, with likelihood ratio tests strongly favoring the power-law over the exponential distribution ($R = 191.085$, $p < 0.001$) and showing no significant difference from the lognormal distribution ($R = -0.040$, $p = 0.859$),

while the simple power-law model is preferred over the truncated power-law ($R = -0.422$, $p = 0.358$). The estimated exponent places our distribution in the critical regime where $2 < \alpha < 3$, implying finite mean but theoretically divergent variance as expressed by $\text{Var}[X] = \int_{x_{\min}}^{\infty} x^2 p(x)dx \propto \int_{x_{\min}}^{\infty} x^{2-\alpha}dx$. While our finite dataset necessarily exhibits finite sample variance, this theoretical property manifests as extreme sensitivity to tail observations and instability in variance estimates, creating fundamental challenges for traditional machine learning approaches that assume stable second moments. The heavy-tailed characteristics create specific methodological vulnerabilities for conventional approaches: (1) assumption violations in methods requiring stable variance, (2) systematic underrepresentation of high-impact tail observations through random sampling, and (3) inability of synthetic oversampling methods to reliably extrapolate extreme tail behavior, which our ensemble approach directly addresses by preserving authentic minority class distribution while ensuring adequate tail representation.

Through bootstrap analysis with 1,000 iterations, we provide empirical validation of extreme loss concentration within the customer base. Our analysis demonstrates that 22.5% of the 9,984 loss-generating customers account for 80% of total losses (95% confidence interval: [20.7%, 24.4%]).

Table 1. Loss Concentration Across Multiple Thresholds

Cumulative Loss	Customer Percentage	95% Confidence Interval
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%]

This concentration effect, validated across multiple thresholds (Table 1 and Figure 1), confirms that a minority of customers drive the majority of losses, justifying our focus on accurate minority class identification rather than overall classification accuracy.

3.2 Binary Classification Framework Under Business Constraints

The binary classification framework operates within strict business implementation constraints that reflect real-world deployment requirements. The insurance company's operational framework allows exclusion of at most 8% of potential customers due to competitive market positioning, regulatory compliance requirements, and customer acquisition targets.This business constraint is non-negotiable and reflects the practical limitations within which any deployed model must operate.

Rather than optimizing this threshold to maximize validation performance, which would risk overfitting to our specific dataset, we accept this business-imposed constraint and set the classification threshold at the 8th percentile of

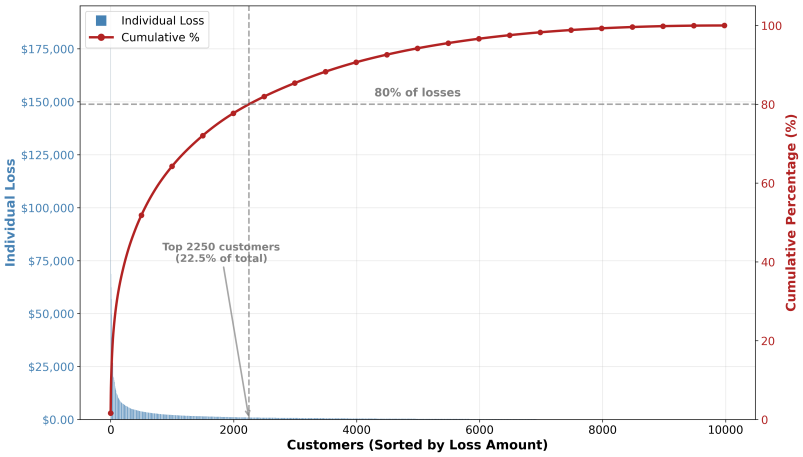

Fig. 1. Cumulative loss distribution of customers

the BMA_corregido distribution, corresponding to values below -€1,708. This approach prioritizes model generalizability and real-world applicability over potentially spurious validation improvements.

The business-constrained threshold creates a severe class imbalance problem with significant practical implications. The resulting binary classification identifies 4,129 customers (8% of the dataset) as high-loss cases, creating a minority-to-majority ratio of approximately 1:11.5.

The minority class, while representing only 8% of customers, accounts for a disproportionate share of negative business impact. Standard machine learning approaches trained on such imbalanced data exhibit systematic bias toward the majority class, often achieving high overall accuracy while failing catastrophically on minority class identification : precisely the opposite of business requirements where minority class detection is highly important.

The asymmetric nature of misclassification costs fundamentally shapes our methodological approach. False negatives (failing to identify high-loss customers) result in direct financial losses, as these customers impose costs far exceeding their premium contributions. Conversely, false positives (incorrectly excluding profitable customers) represent opportunity costs but not direct losses.

Our analysis demonstrates that the 4,129 high-loss customers in our dataset would generate expected losses greatly exceeding the forgone profits from correctly classified profitable customers, validating the business logic of trying to prioritize minority class recall over overall classification accuracy.

3.3 Balanced Bagging Ensemble Approach

Our approach builds upon the established foundation of ensemble methods for class imbalance learning [17], while introducing novel elements specifically designed for heavy-tailed financial risk assessment. While methods such as EasyEnsem-

ble [18] and SMOTEBoost [19] 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. [6], which were implemented under identical conditions.

The theoretical motivation for balanced bagging in heavy-tailed loss scenarios stems from three key principles:

First of all, 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.

The algorithm constructs K base classifiers, where each classifier h_k is trained on a balanced subset $\mathcal{D}_k$ created through controlled sampling:

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 $\mathcal{H} = \{h_1, h_2, \ldots, h_K\}$

```
Partition D into D_maj and D_min based on class labels
for k = 1 to K do
    Sample ⌊|D_min| · r⌋ instances from D_maj to form D_maj^(k)
    Create balanced subset D_k = D_min ∪ D_maj^(k)
    Train base classifier h_k on D_k using XGBoost with fixed hyperparameters
    Add h_k to ensemble H
end for
return H
```

We employ XGBoost [20] 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 first stage constitutes of conducting a preliminary random grid 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 $r = 4.0$, 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 discrete values $\{1.17, 1.47, 2.12, 2.80, 3.12, 3.20, 3.60, 3.85\}$. 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 calculated using non-parametric (bootstrap) methods. Statistical significance was assessed using Friedman tests for multiple comparisons ($\chi^2 = 98.67$, $p < 0.0001$) and Wilcoxon signed-rank tests for pairwise comparisons ($\alpha = 0.05$). The profit function calculates realized benefit by excluding customers classified as high-loss (class 1) and summing the BMA_corregido values for retained customers (class 0):

$$\text{Profit}(r) = \sum_{i \in \mathcal{I}_{retained}(r)} \text{BMA\_corregido}_i$$

where $\mathcal{I}_{retained}(r) = \{i : \hat{y}_i(r) = 0\}$ represents the set of customers retained 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—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.

4 Results

4.1 Sampling Strategy Optimization Results

The grid 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. Grid Search Results for Sampling Strategy Optimization

Sampling Strategy	Mean Real Benefit (€)	Bootstrap 95% CI (€)	Clients Excluded (%)	F1-Score (Mean)	Constraint Compliant
1.17	727,833	[687,621-766,880]	27.6	0.797	No
1.47	713,699	[674,201-754,649]	20.5	0.836	No
2.12	616,093	[561,111-671,588]	12.4	0.874	No
2.80	536,471	[472,676-595,574]	8.2	0.890	No
3.12	505,111	[436,844-566,107]	7.0	0.894	Yes
3.20	489,752	[422,913-553,440]	6.8	0.894	Yes
3.60	460,843	[399,643-526,616]	5.7	0.897	Yes
3.85	435,281	[382,279-491,054]	5.1	0.898	Yes

Statistical significance testing revealed large differences between strategies. The Friedman test confirmed significant variation across all strategies ($\chi^2 = 98.67$, $p < 0.0001$). While sampling ratios 1.17 and 1.47 achieved the highest absolute profits (€727,833 and €713,699 respectively), these strategies violated the operational constraint by excluding 28% and 21% of customers respectively. Among constraint-compliant strategies ($\leq 8\%$ exclusion), sampling ratio $r^* = 3.12$ emerged as optimal, achieving the highest real benefit (€505,111, Bootstrap 95% CI: [436,844-566,107]) while maintaining 7.0% customer exclusion and an F1-score of 0.894. Although sampling ratio 2.80 achieved higher absolute profit (€536,471), it violated the operational constraint with 8.2% customer exclusion, rendering it unsuitable for deployment.

4.2 Model Performance on Validation Set

The optimized balanced ensemble with $r^* = 3.12$ demonstrated robust performance on the validation set comprising 7,227 customers. Figure 2 summarizes the classification performance metrics.

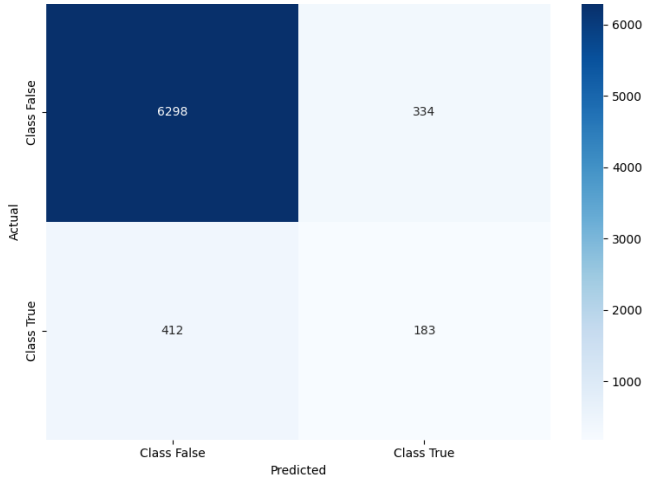

Fig. 2. Confusion matrix of the balanced bagging model for the validation set

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, representing 41% of the theoretical maximum benefit (€1,728,523). This translates to an average benefit of €104.54 per customer while excluding exactly 7.15% of the customer base, well within the operational constraint.

4.3 Final Test Set Evaluation

The final model evaluation on the independent test set of 15,486 customers confirmed the robustness and generalizability of the balanced ensemble approach. Figure 3 presents the comprehensive test set performance metrics.

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 on the test set, representing 36% of the

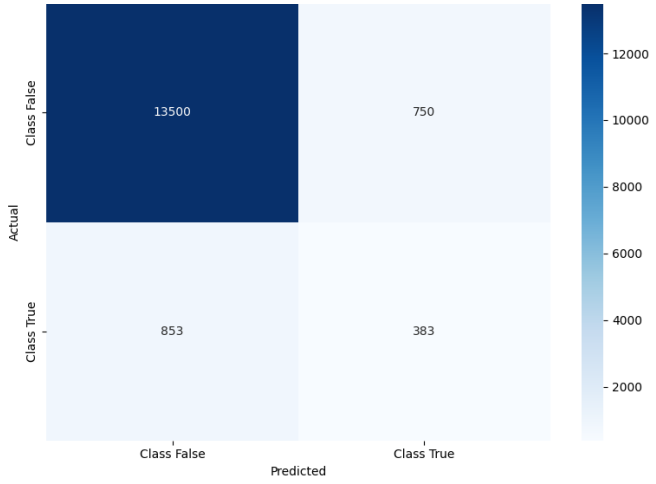

Fig. 3. Confusion matrix of the balanced bagging model for the test set

theoretical maximum benefit (€3,674,846). This corresponds to an average benefit of €92.60 per customer while excluding 7.32% of potential customers, maintaining strict adherence to operational constraints.

4.4 Economic Impact and Misclassification Analysis

The detailed analysis of test set predictions reveals interesting economic trade-offs 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 retained
True Positives	383	-2,606.71	-998,371	Correctly excluded
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 average loss) compared to false positives (€100 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 out of €-

1,976,944), with the single worst case representing a €-176,973 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 excluded customers represent 23.3% of foregone profit (€17,534 out of €75,302), with the maximum individual opportunity cost of €4,393. The relatively modest false positive impact, combined with their distributed nature, suggests that exclusion errors impose manageable opportunity costs compared to retention errors. The model demonstrates strong performance in risk concentration mitigation. True positives exhibit an average loss of €-2,607 per customer, while correctly retained customers (true negatives) show an average profit of €245 per customer. This €2,852 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 successful exclusion of 383 high-risk customers with total expected losses of €998,371.

4.5 Comparative Analysis with Baseline Methodology

To evaluate the effectiveness of our balanced ensemble approach, we compare our results with the methodology proposed by Soriano-Gonzalez et al. [6], 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 baseline methodology from Soriano-Gonzalez et al.

Table 4. Performance Comparison: Balanced Ensemble vs. Baseline Methodology

Metric	Soriano-Gonzalez et al.	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 Exclusion 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 key 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 average) than false positives (€100 average). Both methodologies respect operational constraints, with exclusion rates of 6.0% (baseline) and 7.0% (our approach), both well within the 8% 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 in losses based on the average high-risk customer loss of €-2,318.

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 excluded. At €100 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 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 hyperparameters across all ensemble members, potentially leaving performance gains unrealized. The baseline methodology employed hyperparameter optimization, which may partially explain their superior precision. A fair comparison would require hyperparameter tuning for each ensemble configuration, increasing computational requirements even more.

While we optimized sampling ratios, we used majority voting for final predictions rather than optimizing decision thresholds. The baseline approach's thresh-

old 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 (3.12) was derived from our specific dataset's characteristics. The method's sensitivity to dataset composition, temporal drift, and feature distributions remains unexplored. The approach may require re-optimization for different portfolios, limiting its practical deployability compared to more robust single-model approaches.

5 Conclusion

This study presents a customer classification approach for the insurance industry that combines ensemble learning with a profit-oriented optimization strategy. The proposed method leverages a carefully optimized balanced bagging framework, capable of effectively handling severe class imbalance and complying with operational constraints commonly faced in real-world insurance settings.

The results demonstrate that the model consistently identifies a higher number of high-risk customers, significantly reducing exposure to extreme financial losses while remaining within the predefined exclusion limits. It maintains high overall performance across multiple evaluation stages, offering measurable benefits in terms of both predictive accuracy and economic value. The modeling strategy ensures a solid balance between the accurate detection of risk and the practical demands of deployment, thanks to the integration of authentic data distribution, ensemble majority voting, and a business-constrained objective function optimized for profit.

Furthermore, the robustness of the approach has been validated on independent test sets, confirming its stability and reliability across different data partitions. The overall framework provides a practical and scalable solution for insurers seeking to enhance risk management through data-driven decision-making.

As future work, we propose exploring model variants that reduce computational demands while maintaining predictive quality, as well as extending the approach to other insurance lines or markets to assess its generalization capacity. Additionally, incorporating adaptive mechanisms that respond to changes in data distribution or customer behavior would help sustain long-term effectiveness.

References

1. D Finger, H Albrecher, and L Wilhelmy. On the cost of risk misspecification in insurance pricing. *Japanese Journal of Statistics and Data Science*, 7(2):1111–1153, 2024.
2. Simone Borg Bruun, Christina Lioma, and Maria Maistro. Recommending target actions outside sessions in the data-poor insurance domain. *ACM Transactions on Recommender Systems*, 3(1):1–24, 2024.
3. Changyue Hu, Zhiyu Quan, and Wing Fung Chong. Imbalanced learning for insurance using modified loss functions in tree-based models. *Insurance: Mathematics and Economics*, 106:13–32, 2022.

4. Patricia Carracedo, David Hervás, and Raquel Soriano-Gonzalez. Class imbalance in insurance fraud detection models. *Available at SSRN 4990942*, 2024.
5. Dimitris Bertsimas and Agni Orfanoudaki. Algorithmic insurance. *arXiv preprint arXiv:2106.00839*, 2021.
6. Raquel Soriano-Gonzalez, Veronika Tsertsvadze, Celia Osorio, Noelia Fuster, Angel A Juan, and Elena Perez-Bernabeu. Balancing risk and profit: Predicting the performance of potential new customers in the insurance industry. *Information*, 15(9):546, 2024.
7. Rahul Sahai, Ali Al-Ataby, Sulaf Assi, Manoj Jayabalan, Panagiotis Liatsis, Chong Kim Loy, Abdullah Al-Hamid, Sahar Al-Sudani, Maitham Alamran, and Hoshang Kolivand. Insurance risk prediction using machine learning. In *The international conference on data science and emerging technologies*, pages 419–433. Springer, 2022.
8. Marketa Krupova, Nabil Rachdi, and Quentin Guibert. Explainable boosting machine for predicting claim severity and frequency in car insurance. *arXiv preprint arXiv:2503.21321*, 2025.
9. Esmeralda Brati, Alma Braimllari, and Ardit Gjeçi. Machine learning applications for predicting high-cost claims using insurance data. *Data*, 10(6):90, 2025.
10. Sebastian Baran and Przemysław Rola. Prediction of motor insurance claims occurrence as an imbalanced machine learning problem. *arXiv preprint arXiv:2204.06109*, 2022.
11. Vibhuti Gupta, Julian Broughton, Ange Rukundo, and Lubna Pinky. Learning unbiased risk prediction based algorithms in healthcare: A case study with primary care patients. *Available at SSRN 4984535*.
12. Yu Li, Yi Zhang, Lu Gan, Gengwei Hong, Zimu Zhou, and Qiang Li. Revman: Revenue-aware multi-task online insurance recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 303–310, 2021.
13. OECD. Leveraging technology in insurance to enhance risk assessment and policyholder risk reduction. https://www.oecd.org/publications/leveraging-technology-in-insurance-to-enhance-risk-assessment-and-policyholder-risk-reduction-4844de05-en.htm, 2023.
14. Carlos M Jarque and Anil K Bera. A test for normality of observations and regression residuals. *International Statistical Review*, 55(2):163–172, 1987.
15. Theodore W Anderson and Donald A Darling. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics*, 23(2):193–212, 1952.
16. Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.
17. Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. 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)*, 42(4):463–484, 2012.
18. Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2009.
19. Nitesh V Chawla, Aleksandar Lazarevic, Lawrence O Hall, and Kevin W Bowyer. Smoteboost: Improving prediction of the minority class in boosting. In *European conference on principles of data mining and knowledge discovery*, pages 107–119. Springer, 2003.

20. Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.