risks

MDPI

Article

Advanced Insurance Risk Modeling for Pseudo-New Customers Using Balanced Ensembles and Transformer Architectures

Finn L. Solly [1], Raquel Soriano-Gonzalez [2], Angel A. Juan [2,3,*] and Antoni Guerrero [2,4]

[1] Esade Business School, Ramon Llull University, 59 Torre Blanca Av., 08172 Sant Cugat, Spain
[2] CIGIP—ValgrAI, Universitat Politècnica de València, Ferrandiz-Carbonell, 03802 Alcoy, Spain; rsorgon@epsa.upv.es (R.S.-G.)
[3] Euncet Business School, Universitat Politècnica de Catalunya, Mas Rubial, 08225 Terrassa, Spain
[4] Baobab AI, 28003 Madrid, Spain
* Correspondence: ajuanp@upv.es

Abstract

In insurance portfolios, classifying customers without a prior history at a given company is particularly challenging due to the absence of historical behavior, extreme class imbalance, heavy-tailed loss distributions, and strict operational constraints. Traditional machine learning approaches, including the baseline methodology proposed in previous studies, typically optimize global predictive accuracy and therefore fail to capture business-critical outcomes, especially the identification of high-risk clients. This study extends the existing approach by evaluating two complementary business-aware classification strategies: (i) a balanced bagging ensemble specifically designed to handle class imbalance and maximize expected profit under explicit customer-omission constraints, and (ii) a lightweight Transformer-based architecture capable of learning richer feature representations. Both approaches incorporate the asymmetric financial cost structure of insurance and operate under operational selection limits. The empirical analysis is conducted on a proprietary large-scale auto insurance dataset comprising 51,618 customers and is complemented by validation on nine synthetic datasets to assess robustness. Model performance is evaluated using statistical tests (ANOVA, Friedman, and pair-wise comparisons) together with business-oriented metrics. The results show that both proposed approaches consistently outperform the baseline methodology ($p < 0.001$) in terms of profit, with the ensemble offering a better balance of performance and efficiency, while the Transformer shows stronger robustness and generalization under data perturbations. The balanced ensemble provides the most favourable trade-off between predictive performance, robustness, interpretability, and computational efficiency, making it suitable for deployment in regulated insurance environments, while the Transformer achieves competitive results and exhibits stronger generalization under data perturbations. The proposed approach aligns machine learning with actuarial portfolio optimization by explicitly integrating profit-driven objectives and operational constraints, offering two practical and scalable solutions for risk-based decision-making in real-world insurance settings.

Keywords: insurance risk modeling; customer classification; high-risk customer detection; balanced ensembles; transformer models

Check for updates

Academic Editor: Paolo Giudici

Received: 21 February 2026
Revised: 31 March 2026
Accepted: 13 April 2026
Published: 17 April 2026

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

1. Introduction

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 (Finger et al. 2024). Predicting the performance of 'pseudo-new' customers (those without prior history at our company) is particularly difficult, as external databases might only provide scarce data, and the costs of misclassification are high (Bruun et al. 2024). In this study, pseudo-new customers are individuals who have not previously held a policy with our insurer but may possess publicly available insurance history from standardized national databases such as the Historical Information System of Motor Insurance SINCO in Spain (https://sincostats.tirea.es, accessed on 16 April 2026). 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 (Hu et al. 2022). In this study, the problem is explicitly defined as a binary classification task under business constraints, where pseudo-new customers are assigned to either a high-risk or a low-risk class. While the model produces probability scores, these are only used for thresholding under business constraints. Still, our final objective is categorical classification rather than continuous risk estimation. Conventional approaches tend to favor the majority class, resulting in poor identification of high-loss customers (Singla et al. 2021). Business constraints add to the challenge, since insurers cannot omit large numbers of customers without risking business volume and market share (Ngwenduna and et al. 2021).

Most existing works address these issues through feature selection and model optimization. Soriano-Gonzalez et al. (2024), for example, predicted pseudo-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 (Uddin et al. 2023)—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 (Marhavilas and Koulouriotis 2021).

To overcome these limitations, this study evaluates two complementary business-aware methodologies: (i) a balanced bagging ensemble to enforce class balance and optimize profit under customer-omission constraints, and (ii) a Transformer-based classifier capable of learning richer feature representations and profiting from augmented training data. The balanced ensemble approach preserves all high-risk customers while systematically undersampling the majority class according to a profit-optimized sampling ratio. Multiple XGBoost base learners are trained on these balanced subsets, and their outputs are aggregated to improve robustness, align predictions with asymmetric costs, and comply with the maximum 8% customer-omission constraint. In contrast, the Transformer model adopts a strategy based on deep representation learning. During training, the model increases data diversity through a predefined augmentation ratio and selects at each epoch the checkpoint that maximizes validation profit. Final decisions for pseudo-new customers are obtained by applying probability thresholds that comply with the operational constraints of the business. These two approaches—one focused on structured imbalance handling and the other on deep representation learning—offer distinct yet complementary solutions to the challenge of assessing the risk of pseudo-new customers in insurance portfolios.

This paper makes three main contributions: First, it formulates the classification of pseudo-new insurance customers as a portfolio value optimization problem under an explicit operational constraint on the maximum admissible omission rate, aligning the learning objective with real underwriting practice. The task involves extreme class imbalance, with roughly 1 high-risk customer for every 11 low-risk customers, and must respect a strict operational constraint of omitting at most 8% of the portfolio. Second, it proposes an imbalance-aware approach that integrates asymmetric economic costs, heavy-tailed loss be-

havior, and profit-driven model selection within a unified decision-making scheme. Third, it provides a comparative evaluation between an interpretable balanced ensemble and a lightweight Transformer architecture under identical business constraints, analyzing their economic performance, robustness to data perturbations, and computational requirements.

The remainder of the paper is organized as follows: Section 2 reviews related work. Section 3 describes the data preparation process, including preprocessing, feature selection, and statistical analysis. Section 4 outlines the proposed pipeline after data curation, incorporating business constraints and the balanced bagging ensemble approach. Section 5 presents the experimental results. Section 6 discusses the findings from both a business and methodological perspective. Finally, Section 7 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 (Fitriani and Febrianto 2021; Leo et al. 2019; Simester et al. 2020). 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 (Hutagaol and Mauritsius 2020; Sahai et al. 2022). 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 (Henckaerts et al. 2021). 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 (Brati et al. 2025; Dietterich 2000; Zhou 2012). Techniques such as XGBoost and explainable boosting machines have been applied to maintain both high accuracy and model transparency (Krupova et al. 2025). For instance, Soriano-Gonzalez et al. (2024) 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. These approaches are typically evaluated using statistical performance metrics, with limited consideration of portfolio-level economic value or operational deployment constraints.

Several studies have focused on imbalanced data problems, as events like fraud or severe claims are rare in relation to the broader customer base (Hanafy and et al. 2021). To address this, researchers have applied undersampling, oversampling (e.g., SMOTE, as in Chawla et al. (2003)), and more recent GAN-based oversampling approaches (Mienye and Swart 2024), as well as class weighting and adaptive ensemble methods. Beyond these general strategies, ensemble-based imbalance methods have also been proposed, including SMOTEBoost (Chawla et al. 2003), RUSBoost (Seiffert et al. 2009), Balanced Random Forest (Pes 2021), and EasyEnsemble2 (Liu et al. 2008). These approaches combine boosting or bagging with resampling to improve minority detection and have been widely adopted in imbalanced learning tasks. However, they do not incorporate explicit business constraints or profit-driven objectives, which are central to our framework.

Baran and Rola (2022) and Khamesian et al. (2022) 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 (Gupta et al. 2025). Nonetheless, many studies optimize mainly for statistical measures such as AUC, which reflects ranking performance rather than direct classification quality, or F1-score, which is more appropriate under class imbalance. However, these metrics often ignore business constraints such as 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, such as online recommendation systems, have applied reinforcement learning and multitask neural networks to optimize conversion rates (Li et al. 2021), 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 (Loisel et al. 2021). 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.

Insurance customers can be broadly categorized according to the type of coverage they hold, each exhibiting distinct risk profiles that necessitate tailored assessment approaches. In life insurance, risk is primarily determined by demographic and socioeconomic factors such as age, medical history, occupation, and asset holdings, which jointly drive risk aversion and additional coverage demand (Lim et al. 2023). In health insurance, risk concentration is driven by utilization patterns and chronic disease prevalence, with a small proportion of high-cost patients accounting for the majority of expenditure. Tree-based ML methods such as Random Forest and logistic regression have demonstrated near-perfect discriminative ability in identifying high-cost utilizers from administrative claims data (Seyam 2025). In auto insurance, risk is jointly determined by driver behavior, vehicle attributes, and contextual factors such as speed limits, weather conditions, and road infrastructure, with tree-based ML methods such as XGBoost and Random Forest demonstrating strong predictive performance in telematics-based risk assessment (Masello et al. 2023). In addition to individual loss concentration, recent work on financial systems has highlighted that rare events (such as large claims or defaults) often propagate through interconnected portfolios via contagion effects. For instance, Giudici and Parisi (2018) discuss how sovereign credit risk spreads are influenced not only by idiosyncratic factors but also by contemporaneous contagion among countries. This network-based propagation suggests that customer risk should be evaluated not in isolation, but within a broader context of interdependencies.

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 (Sadreddini et al. 2021). Nevertheless, classification of pseudo-new customers with scarce historical information remains limited, especially when data acquisition is costly or incomplete (Sari and Purwadinata 2019). In such settings, the financial impact of misclassification can be high, but only a limited number of studies explicitly address this problem from a business-oriented perspective (Tian et al. 2023). 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 (De Meulemeester and De Moor 2020; Elbhrawy et al. 2024; LeCun et al. 2015). Natural language processing has also been integrated with machine learning to analyze text from policy documents, claims, and customer interactions (Cambria and White 2014; Kolambe and Kaur 2023; Young et al. 2018). 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 (Doshi-Velez and Kim 2017; Orji and Ukwandu 2024). Tools such as SHAP (Shapley additive explanations) and LIME have been adopted to improve the transparency of model decisions (Le et al. 2023; Lundberg and Lee 2017). Still, SHAP values provide estimates of feature importance, and their reliability can decrease in high-dimensional settings, where heuristics are often used in practice instead of exact attribution. In parallel, feature selection strategies tailored to insurance, such as ordinal risk categories, have also been explored (Ghorbani and Soriano-Gonzalez 2024). Other ongoing concerns are data privacy and compliance with regulations such as the General Data Protection Regulation in the European Union (Olasehinde et al. 2025; Voigt and Von dem Bussche 2017). These constraints must be considered in the practical deployment of machine learning models in insurance workflows.

Recent advances have applied deep learning and feature-interaction models to tabular insurance data. Transformer-based models (e.g., TabTransformer, FT-Transformer) capture complex dependencies across features (Gorishniy et al. 2021), and TabNet has shown competitive results on public benchmarks (Shah et al. 2022). Nonetheless, these methods typically require substantial computational resources and offer limited interpretability, which is a key requirement in regulated insurance contexts. In practice, post hoc attribution can be computationally intensive and is often applied outside the training loop. Moreover, while deep learning models can achieve strong predictive performance, their application in insurance remains constrained by transparency requirements, regulatory obligations, and the operational need for models that can be audited and explained. As a result, research has increasingly focused on balancing predictive accuracy with interpretability, motivating continued exploration of architectures capable of modeling complex tabular interactions while remaining suitable for real-world deployment (Black and Murray 2019; Henckaerts et al. 2021).

These models are widely used in life insurance due to their ability to model customer risk profiles, handle heterogeneous data, and provide interpretable or flexible predictive structures. Traditional methods such as logistic regression and decision trees offer interpretability, which is essential for actuarial decision-making, while ensemble methods and support vector machines improve predictive performance in complex scenarios. Although much of the literature focuses on life insurance, these models are also applicable to non-life insurance contexts, including motor insurance, where similar challenges arise, such as class imbalance, heterogeneous feature spaces, and the need to model claim risk and customer profitability. Therefore, these models provide a relevant methodological foundation for the present study.

To provide a structured overview, Table 1 summarizes the main state-of-the-art (SOTA) techniques most frequently applied to imbalance and tabular learning problems, highlighting their strengths, limitations in insurance contexts, and how our proposed approach differs. This comparative perspective emphasizes that, unlike existing methods, our framework directly integrates business constraints and profit-driven optimization while maintaining interpretability and practical feasibility.

We adopt a balanced bagging ensemble with profit-driven optimization because, unlike synthetic oversampling or purely cost-sensitive methods, this approach preserves the authenticity of minority cases, directly incorporates business constraints ($x \leq 8\%$ omission), and provides interpretable results suitable for regulated insurance contexts. Additionally, we include a Transformer-based model to complement the ensemble approach, allowing us to benchmark a modern deep learning architecture within the same business-constrained framework. Rather than replacing traditional methods,

the Transformer serves as a high-capacity alternative whose performance can be contrasted against an interpretable ensemble solution. This enables a broader evaluation of methodological trade-offs, balancing interpretability and representational power when addressing pseudo-new customer risk classification in insurance.

Table 1. Comparison of SOTA imbalance methods with the proposed approach.

Method	Strengths	Limitations	Difference vs. Our Approach
SMOTE/GAN oversampling	Increases minority samples via synthesis (Chawla et al. 2003; Mienye and Swart 2024).	May create unrealistic cases; poor fit for heavy-tailed data.	Keeps only real cases, no synthetic artifacts.
Cost-sensitive/Focal Loss	Penalizes costly misclassifications; improves recall (Sari and Purwadinata 2019; Tian et al. 2023).	Ignores business limits (e.g., $\leq$8% omission).	Improves recall but does not explicitly enforce operational constraints or optimize portfolio-level profit.
TabNet (deep learning)	Captures feature interactions; strong benchmarks (Shah et al. 2022).	High cost; low interpretability; limited regulatory use.	Interpretable tree ensembles, easy deployment.
Transformer models	Models complex, high-order dependencies (Gorishniy et al. 2021).	Requires large numbers of data; expensive; low transparency.	Provides high representational capacity but does not explicitly incorporate profit-driven objectives or operational constraints as in our approach.
SHAP-based selection	Non-linear feature importance; interpretability (Le et al. 2023; Lundberg and Lee 2017).	Heavy computation; post hoc only.	Mutual information used within training pipeline.
SMOTEBoost/RUSBoost/ Balanced RF/EasyEnsemble2	Combines boosting/bagging with resampling; widely used (Chawla et al. 2003; Liu et al. 2008; Pes 2021; Seiffert et al. 2009).	Does not preserve real minority; lacks profit/business focus.	Preserves genuine minority, profit-optimized, business limits enforced.
Proposed ensemble	Preserves minority; profit-driven; interpretable.	Slightly lower precision; higher cost than single models.	Directly aligned with insurance business and regulation.

To the best of our knowledge, the joint consideration of operational rejection constraints, profit-driven model selection, and imbalance learning under heavy-tailed losses has received limited attention in the context of pseudo-new customer underwriting. By addressing these three dimensions within a unified framework, this study moves the modeling objective from pure statistical accuracy toward actuarial portfolio value optimization under realistic business conditions.

3. Data Preparation

This section describes the data preparation process required to ensure that the predictive models are trained under the same information constraints faced at the customer acquisition stage.

The empirical analysis uses a real-world insurance dataset with 116,934 customer records covering 2016 to 2023. The raw dataset initially contained 247 features on demographics, vehicle attributes, policy details, and historical claims. To ensure applicability to prospective customers, we removed variables not available at acquisition (e.g., post-policy information and claim outcomes).

3.1. Statistical Analysis

We tested the distributional properties of the continuous target variable, defined as the customer's average annual profit (AAP). The customer's annual profit is defined as the difference between the premium paid and the claims cost incurred during the same period, i.e., profit = premium $-$ claims. Positive values indicate profitable customers, while negative values correspond to loss-generating customers. For clarity, we use the term "profit" consistently throughout the paper to refer to this metric. The null hypothesis assumes

that the annual profit follows a standard probability distribution (normal, exponential, logistic, or Gumbel). The Jarque–Bera test (Jarque and Bera 1987) (1.81×10^8, $p < 0.001$) rejects normality, and Anderson–Darling tests (Anderson and Darling 1952) reject normal, exponential, logistic, and Gumbel probability distributions at all significance levels. Losses exhibit a heavy-tailed distribution, concentrating a large proportion of the portfolio risk in a small number of customers. Of 51,618 customers, 9984 (19.3%) generated negative returns, with total losses of 11,328,603 euros, while the portfolio as a whole produced only 909,881 euros in net profit before intervention. These figures result from aggregating individual returns: the cumulative claims of loss-generating customers exceeded their premiums by 11.3 million euros, whereas, across the full portfolio, premiums only outweighed claims by 0.9 million euros. Thus, loss-generating customers imposed losses of more than 12 times the net profit, underscoring the need for accurate identification.

Following Clauset et al. (2009), we conducted a power-law analysis of the tail behavior of customer annual losses, defined as negative values of the annual profit distribution. The maximum-likelihood estimator yielded an exponent $\alpha = 2.381 \pm 0.057$ with $x_{\min} = 3454$ euros. 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 log-normal 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 the extreme sensitivity to tail values and unstable variance estimates. These features can negatively affect conventional machine learning methods, 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. The proposed approaches, both the ensemble and the Transformer, address these limitations by maintaining a consistent representation of extreme cases in the data. This is particularly relevant from an actuarial perspective, where portfolio performance is driven by a small number of high-severity observations.

Moreover, bootstrap analysis with replacement using 1000 iterations confirmed extreme loss concentration: 22.5% of loss-generating customers accounted for 80% of total losses, with a 95% confidence interval (CI) of (20.7%, 24.4%). The bootstrap procedure used simple random sampling with replacement from the original dataset, with each bootstrap sample maintaining the original dataset size ($n = 9984$). As shown in Table 2 and Figure 1, a minority of customers drive most losses, reinforcing the need to prioritize minority class identification over global classification accuracy.

Table 2. 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%)

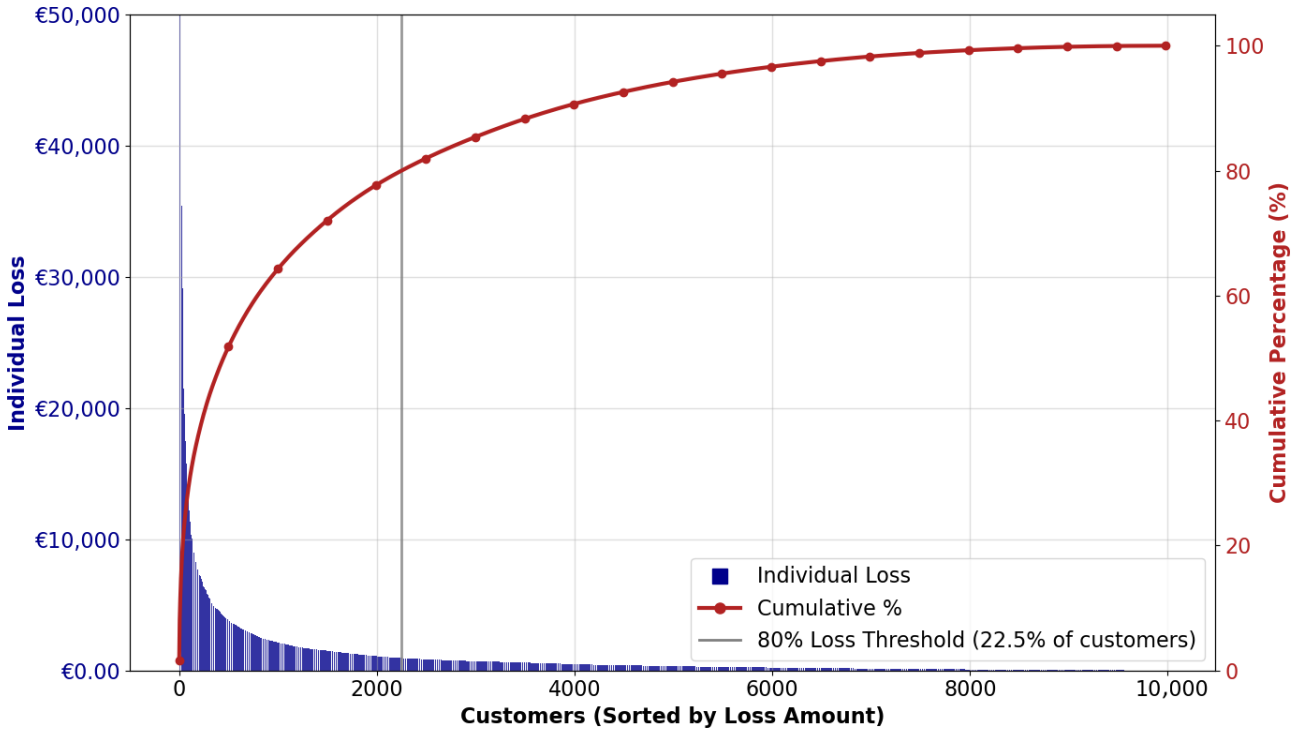

Figure 1. Cumulative loss distribution. Blue bars show individual losses, the red curve the cumulative percentage. The top 22.5% of customers account for 80% of total losses.

3.2. Preprocessing and Feature Selection

Following Soriano-Gonzalez et al. (2024), we applied a standard preprocessing pipeline that included one-hot encoding of categorical variables and the standardization of column names (e.g., converting names to lowercase, removing special characters, and ensuring consistent naming conventions) to ensure computational compatibility. Missing values were imputed using SoftImputer, implemented in the Python `fancyimpute` (version 0.7.0) package, which applies iterative matrix completion with nuclear norm regularization.

The dataset includes a heterogeneous set of variables comprising continuous, categorical, and count-based features. Continuous variables correspond to numerical attributes such as vehicle value, premium amounts, age, and aggregated socio-demographic indicators. Categorical variables include encoded identifiers related to geographic location (e.g., region and municipality), prior insurance company, and customer profile characteristics such as credit capacity. They also include multiple binary indicators derived from one-hot encoding, representing product characteristics such as the type of insured vehicle and selected coverage options. Additionally, the dataset contains discrete count variables related to claims history (e.g., number and frequency of claims in the national insurance database), which provides relevant information on customer risk profiles.

For the ensemble method, feature selection was performed using mutual information, which measures the amount of information shared between each input variable and the target, capturing both linear and non-linear dependencies. This criterion allows identification of the most informative features for predicting high-risk customers (Tiwari et al. 2024).

Customers lacking data in SINCO, which provides access to Spain's Historical Automobile Insurance Information System (SIHSA), were omitted, leaving 51,618 observations with 196 features. This exclusion was necessary because SINCO variables provide the only standardized information on prior insurance history and claims, which are critical predictors of customer risk. Without these data, the model would face severe information loss and potential bias in risk classification; therefore, we retained only customers with

complete SINCO records. Although this filtering reduced the sample size, it reflects a realistic underwriting scenario in which standardized historical information is required for risk assessment.

For the ensemble method, feature selection was performed using mutual information instead of correlation analysis. Mutual information scores were computed for all 194 inputs with respect to the binary classification target, and the 100 features with the highest scores were retained based on their predictive relevance for identifying high-risk customers. This process removed 53 features with very little predictive information (MI scores < 0.001). In contrast to correlation-based approaches, mutual information accounts for non-linear relationships and links feature selection directly to the classification goal while maintaining interpretability. This also reduces model complexity and facilitates periodic retraining in operational environments. However, for the Transformer model, applying feature selection resulted in a notable drop in performance, so all 194 input features were used.

The dataset was partitioned into three independent subsets using stratified sampling to preserve representative class distributions: training (56%), validation (14%), and test (30%). For the ensemble models, the training set was used to fit models under different sampling ratio configurations, and the validation set was used to get the optimal sampling ratio r via profit-based evaluation. For the Transformer model, the training set is used for parameter fitting, while the validation set serves to monitor and control overfitting during training, and to select the best-performing model in the training phase based on validation profit. The test set was kept completely separate and used only for the final unbiased assessment of the selected configuration in both approaches. This design prevents information leakage and ensures an unbiased estimation of economic performance.

4. Modeling and Analysis Pipeline

4.1. Balanced Ensemble Model

4.1.1. Binary Classification Under Business Constraints

The binary classification task was formulated as a portfolio selection problem 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. Rather than tuning this threshold to maximize validation metrics, which risks overfitting, we adopted the business-imposed constraint directly. The classification threshold was set at the 8th percentile of the target variable (customer's average annual profit), corresponding to values below −1708 euros, emphasizing generalizability over dataset-specific performance gains. This mirrors real underwriting practice, where acceptance limits are externally defined and cannot be tuned to optimize model performance.

This constraint produces severe class imbalance with direct implications for model choice. We flagged 4129 customers (8% of the dataset) as high-loss cases, yielding a minority-to-majority ratio of 1:11.6. Traditional classifiers achieved high overall accuracy (85–92%) but failed to identify high-risk customers. Random Forest achieved 91.7% accuracy yet detected only 0.2% of high-risk customers (1 out of 595 cases). Logistic regression showed similar performance with 91.6% accuracy but only 0.5% recall for the minority class. These results show that accuracy alone can be misleading, hiding poor performance on critical minority classes.

The 1:11.6 imbalance ratio creates three key problems: First, algorithms are overwhelmed by majority class patterns; for instance, a dummy classifier predicting only the majority class achieved 91.8% accuracy while never identifying a single high-risk customer. Second, even sophisticated methods struggle to balance precision and recall. Both Random Forest and logistic regression achieved reasonable precision (20%) but poor recall (<1%) for

high-risk cases. Third, these detection failures have severe business consequences: standard classifiers captured only 8.5–13.8% of maximum possible profit, compared to 27.2% with improved minority class detection.

The costs of misclassification are highly asymmetric. Missing high-loss customers (false negatives) leads to direct financial losses, while incorrectly flagging profitable customers (false positives) creates opportunity costs but no direct losses. This cost asymmetry prioritizes minority class recall over overall accuracy because portfolio profitability is primarily driven by the correct identification of high-severity customers rather than by global classification performance.

4.1.2. Balanced Bagging Ensemble Approach

Our approach builds upon ensemble methods for class imbalance learning (Galar et al. 2012), introducing novel elements for heavy-tailed financial risk assessment. While methods such as EasyEnsemble (Liu et al. 2008) and SMOTEBoost (Chawla et al. 2003) have demonstrated effectiveness in general imbalanced scenarios, the specific requirements of financial risk assessment—extreme loss concentration, business constraints, and asymmetric costs—motivate our specialized approach. In this context, the objective is not only predictive performance but the maximization of portfolio economic value under operational constraints.

The theoretical motivation for balanced bagging in heavy-tailed loss scenarios stems from three principles: First, in power-law distributed losses ($\alpha = 2.381$), accurate minority class identification yields disproportionate business value; 22.6% of loss-generating customers account for 80% of total losses. Second, unlike synthetic oversampling, our approach maintains genuine characteristics of minority class instances, which is crucial for heavy-tailed distributions where extreme tail behavior cannot be reliably synthesized without introducing distributional artifacts. Third, by training each base learner on different majority class samples while maintaining complete coverage of the minority class, we ensure thorough exploration of the majority class decision space with consistent reinforcement of the minority class.

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 (Chen and Guestrin 2016) as the base learner due to its demonstrated effectiveness in handling non-linear relationships, feature interactions, and gradient-based optimization. Although XGBoost is itself an ensemble method, we use it as a base learner for several reasons: (i) each XGBoost model sees a different balanced sample of the data, creating diversity at the data level rather than just the model level; (ii) XGBoost's gradient boosting focuses on sequential error correction, while our bagging approach provides parallel diversity through different data distributions; and (iii) the combination addresses both algorithmic bias (through boosting) and sampling bias (through balanced bagging).

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 $\mathcal{H} = \{h_1, h_2, \ldots, h_K\}$
1: Partition $\mathcal{D}$ into $\mathcal{D}_{\text{maj}}$ and $\mathcal{D}_{\text{min}}$ based on class labels
2: **for** $k = 1$ to K **do**
3: Sample $\lfloor |\mathcal{D}_{\text{min}}| \cdot r \rfloor$ instances from $\mathcal{D}_{\text{maj}}$ to form $\mathcal{D}_{\text{maj}}^{(k)}$
4: Create balanced subset $\mathcal{D}_k = \mathcal{D}_{\text{min}} \cup \mathcal{D}_{\text{maj}}^{(k)}$
5: Train base classifier h_k on $\mathcal{D}_k$ using XGBoost with fixed hyperparameters
6: Add h_k to ensemble $\mathcal{H}$
7: **end for**
8: **return** $\mathcal{H}$

The optimal sampling ratio r was determined through a two-stage optimization process. First, a preliminary random search across $r \in [1.0, 20.0]$ identified promising regions, revealing that performance gains decrease significantly beyond $r = 4$ while computational costs increase linearly. Second, a systematic grid search across $r \in \{1.17, 1.47, 2.12, 2.80, 3.12, 3.20, 3.60, 3.85\}$ was performed, with each configuration evaluated on the same validation set to ensure consistent comparison. Let $\hat{y}_i(r)$ denote the predicted class for customer i under sampling ratio r, where $\hat{y}_i(r) = 1$ indicates exclusion (high risk) and $\hat{y}_i(r) = 0$ indicates acceptance (low risk). The optimization framework employs constrained profit maximization as follows:

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

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

The constraint ensures that the proportion of customers flagged for exclusion does not exceed 8% of the validation set. This constraint applies to the total exclusion rate (both correctly identified high-risk customers and incorrectly flagged profitable customers). The enforcement of this constraint at test time is described in Section 5.1.2. 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 pair-wise comparisons ($\alpha = 0.05$). In addition to profit-based evaluation, standard classification metrics are reported for completeness, including the weighted F1-score, which is computed as the average of the class-wise F1-scores weighted by the number of instances in each class. The profit function calculates realized profit by omitting customers classified as high-loss (class 1) and summing the values of average annual profit for selected customers (class 0):

$$\text{Profit}(r) = \sum_{i \in \mathcal{I}_{selected}(r)} \text{AAP}_i$$

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. This formulation is equivalent to selecting the subset of customers that maximizes expected portfolio profit subject to an acceptance capacity constraint.

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

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

where $1[\cdot]$ denotes the indicator function. This aggregation strategy provides a measure of prediction stability through the vote distribution and enhances robustness against individual classifier errors. 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. In addition, the use of tree-based base learners preserved model transparency, allowing standard feature-importance analysis for underwriting interpretation.

4.2. Transformer-Based Model

The proposed deep learning model is based on a standard Transformer encoder architecture. It is composed of 3 layers, where each block includes a multi-head self-attention mechanism with 8 heads and a position-wise feedforward network using ReLU activations, with an embedding dimension of 128 and a feedforward dimension of 512. The input data, consisting of 194 features, are first projected into a 128-dimensional space through a linear layer. Instead of treating each feature as a separate vector and adding positional encodings, all inputs are combined into a single embedding vector. This design choice was implemented for computational efficiency, as separating the inputs significantly increased training time. These embeddings are then processed by the Transformer encoder, and the output is passed through another linear layer that performs the final classification. This classification layer also includes a dropout mechanism, intended both to reduce overfitting during training and to improve the model's generalization to unseen data.

Moreover, the Transformer does not output binary labels (0 for good clients and 1 for potential high-loss customers) but rather continuous values in the range $(-\infty, +\infty)$. Clients with scores below zero are considered good, whereas those with scores above zero are classified as potentially bad. During training, the binary cross-entropy loss is directly used on these raw logits for numerical stability. While clients are directly evaluated in this logit space at the inference time, a sigmoid function can optionally be applied to map the scores into a $(0, 1)$ range if one needs to interpret them as probabilities. To satisfy the business constraint that no more than 8% of clients can be rejected, the model output is treated as a ranking, and the 8% of clients with the highest scores are thus identified as rejected cases. This architecture is simpler than other Transformer-based models for tabular data, such as the FT-Transformer introduced by Gorishniy et al. (2021). However, as explained in previous sections, training speed was a key requirement for the company, motivating the use of a lightweight model. This design reflects a realistic trade-off between representational capacity and training time in production environments. The architecuture used can be seen in Figure 2.

For the training data, the class imbalance was addressed by oversampling instances of potentially bad clients. Specifically, the training set was augmented by duplicating existing samples of bad clients until a ratio of 3.5 between good and bad clients was achieved. This procedure ensured that the model was exposed to a sufficient number of high-risk examples during training, and the value of 3.5 was determined by evaluating ratios from 2.0 to 4.0 in steps of 0.1, using the validation profit as the selection criterion. The final model was chosen based strictly on its performance on a separate validation set of unseen data. This ensured the model actually learned to generalize instead of just memorizing the repeated samples, preventing overfitting. The model was then trained for 25 epochs with a learning rate of 5×10^{-5} using the Adam optimizer (Kinga et al. 2015), and, after each epoch, the profit on the validation set was computed. Among all epochs, the model that achieved the highest validation profit was selected, and it was subsequently used to generate the final results.

Methodology Overview

To improve readability, Figure 3 provides a visual summary of the proposed methodology. It shows the main steps of the process, from data preparation and feature selection to the application of business constraints, training of the balanced bagging ensemble, and final decision-making through majority voting. In addition, the diagram incorporates the Transformer-based model as a parallel workflow, illustrating how both approaches share the same data preparation pipeline but diverge into two distinct modeling strategies—one based on balanced bagging and one based on deep representation learning. This diagram

complements the pseudo-code (Algorithm 1), the Transformer training description, and the mathematical formulation by providing a unified visual overview of the complete workflow for both models. This unified view highlights the integration of statistical modeling, economic evaluation, and operational constraints within a single decision-making framework.

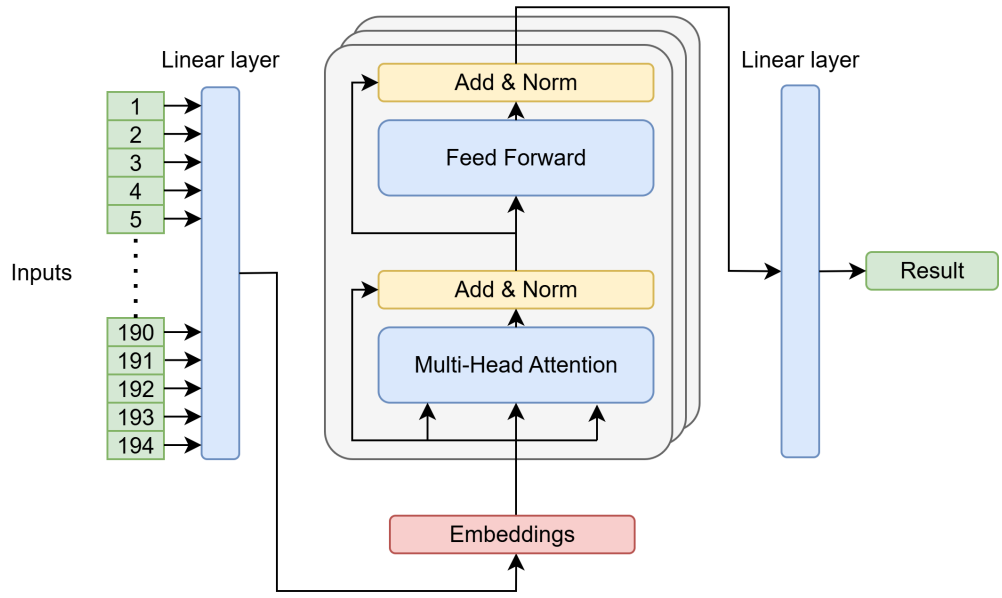

Figure 2. Architecture of the Transformer approach.

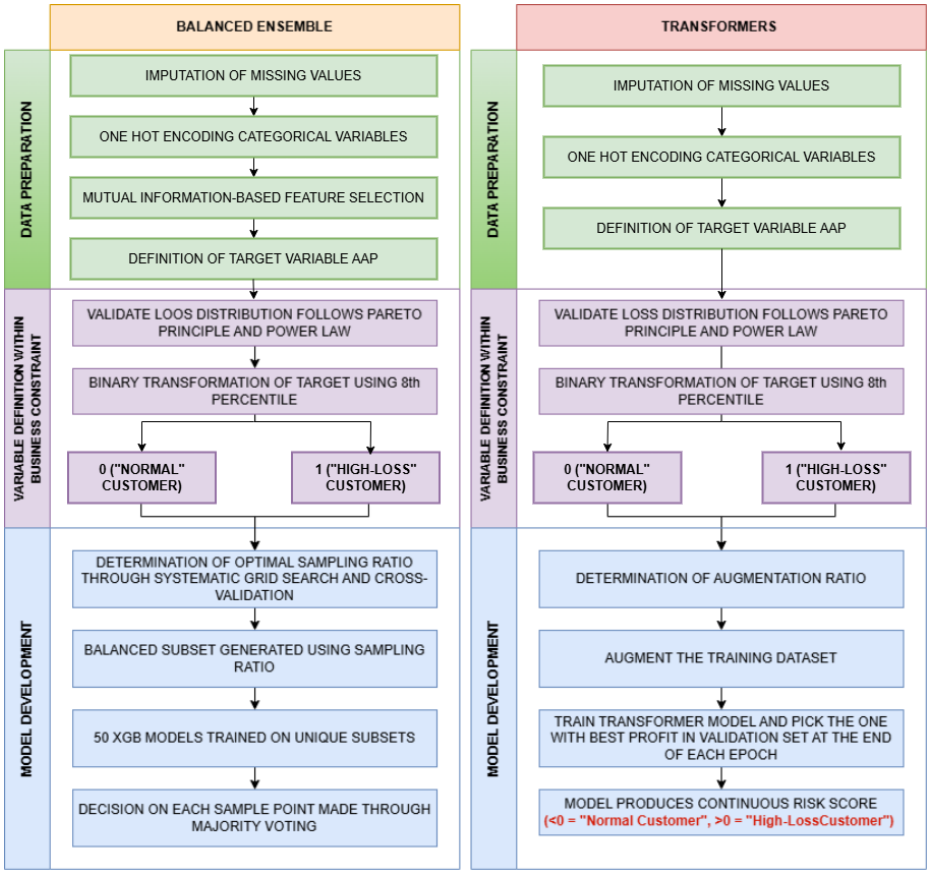

Figure 3. Diagram of the proposed methodology illustrating the shared data preparation pipeline and the two modeling approaches developed in parallel: the balanced bagging ensemble and the Transformer-based model.

5. Results

This section reports the empirical performance of both approaches under the same operational constraint of omitting at most 8% of customers. In addition to standard classification metrics, we emphasize portfolio-level economic outcomes, since the objective is value optimization under realistic underwriting limits.

5.1. Results for the Balanced Ensemble Model

5.1.1. Sampling Strategy Optimization Results

The sampling ratio optimization evaluated eight candidate values in the range $r \in [1.17, 3.85]$, with each configuration assessed on the same hold-out validation set to ensure consistent comparison.

Table 3 summarizes the sampling ratio optimization results. The Friedman test confirmed significant variation across 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 euros), they violated the constraint by omitting 28% and 21% of customers. This highlights the practical trade-off between unconstrained profit maximization and operational feasibility under underwriting capacity limits. Among compliant strategies ($\leq$8% omission), $r^* = 3.12$ achieved the highest profit (505,111 euros, 95% CI: (436,844, 566,107)), with 7.0% omission and an F1-score of 0.894. Sampling ratio 2.80, despite achieving a higher profit (536,471 euros), violated the constraint at 8.2% omission. With 50 models each sampling 7172 majority instances from a pool of 26,606, all majority class instances appeared at least once (average frequency: 13.5 times per instance).

Table 3. Sampling ratio optimization results on the validation set.

Sampling Strategy	Mean Profit (Euros)	Bootstrap 95% CI (Euros)	Omission Rate (%)	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

5.1.2. Model Performance on the Test Set

The optimized ensemble with $r = 3.12$ was evaluated on the test set (15,486 customers). Figure 4 and Table 4 summarize the performance metrics. The model achieved high performance on the majority class (F1: 0.94) but limited minority class recall (30%, F1: 0.32), reflecting the inherent challenge of detecting high-risk customers under severe imbalance. From a portfolio perspective, this recall level translated into a substantial reduction in tail-driven losses while respecting the 8% omission constraint.

The confusion matrix revealed 13,463 true negatives, 376 true positives, 787 false positives, and 860 false negatives. Performance metrics remained consistent with validation results. The model achieved a real profit of 1,459,411 euros (40% of maximum), corresponding to 101.89 euros per customer, while omitting 7.51% of customers. This omission rate confirms that the constraint was satisfied at the test time; rather than applying a hard rank-based cutoff, the ensemble enforced the constraint implicitly through the choice of r^*, which was selected on the validation set to produce compliant omission rates, ensuring that majority voting at the test time naturally yielded a compliant result.

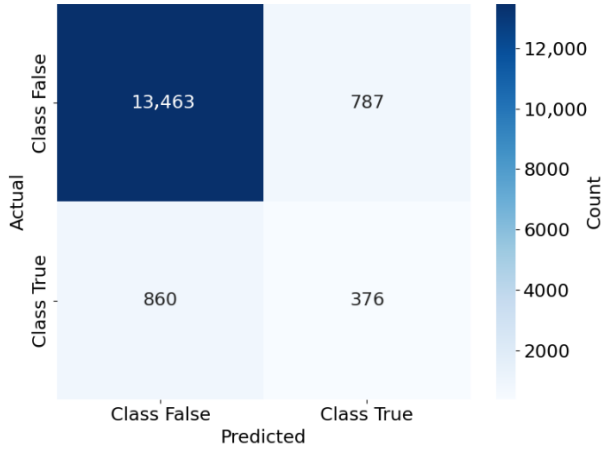

Figure 4. Confusion matrix of the balanced bagging model on the test set, showing 31% recall for high-risk customers.

Table 4. Classification report of the balanced bagging model on the test set.

Class	Precision	Recall	F1-Score	Support
False	0.94	0.94	0.94	14,250
True	0.32	0.30	0.31	1236
Accuracy			0.89	15,486
Macro avg	0.63	0.62	0.63	15,486
Weighted avg	0.89	0.89	0.89	15,486

5.1.3. Economic Impact for the Ensemble Model

The detailed analysis of test set predictions reveals economic trade-offs that are inherent in customer classification under business constraints. Table 5 presents the comprehensive breakdown of misclassification costs and their business implications.

Table 5. Economic impact analysis by customer category for the ensemble model.

Category	Count	Mean Profit (eu)	Total (eu)	Business Implication
True Negatives	13,463	245.40	+3,303,887	Correctly selected
True Positives	376	−3007.55	−1,130,838	Correctly omitted
False Negatives	860	−2144.74	−1,844,476	Missed high-risk
False Positives	787	98.39	+77,434	Foregone profit

False negatives impose higher per-customer costs (−2145 euros) than false positives (98 euros), validating the recall-prioritized approach. The ten worst false negatives accounted for 24.4% of total false negative losses (−481,889 euros), with the worst case at −62,398 euros, demonstrating that extreme cases remain challenging despite overall good performance. False positives showed a more distributed impact: the ten highest-value omitted customers represented 23.3% of foregone profit (17,534 euros), with a maximum individual cost of 4393 euros. True positives averaged −3008 euros loss per customer, while true negatives averaged 245 euros profit, yielding a 3253 euro differential. False negatives were concentrated in the −1000 to −2000 range, indicating systematic challenges in distinguishing moderate-loss customers from profitable customers.

5.2. Results for the Transformer-Based Model

5.2.1. Augmentation Strategy Optimization Results

For the Transformer approach, the augmentation ratio did not affect compliance with the 8% omission constraint since the model outputs a continuous ranking and always

removes the top 8% highest-risk clients. Consequently, the sampling ratio was tuned only with respect to validation profit. A search was performed over ratios between 2.8 and 4.0, using increments of 0.1, and each value was evaluated using the same validation set to ensure a consistent comparison across configurations. For every ratio, the model was trained and the corresponding validation profit was recorded, allowing a direct assessment of the effect of each oversampling level on performance. This procedure indicated that a ratio of 3.5 achieved the highest validation profit, and this value was therefore selected for training the final model. Values outside the range of 2.8 to 4.0 produced poor performance, so the search was concentrated within this interval.

5.2.2. Model Performance on Test Set

The Transformer model trained with an augmentation ratio of $r = 3.5$ was evaluated on the test set (15,486 customers), and Table 6 summarizes the resulting classification performance. The model performed strongly on the majority class (F1-score: 0.94), which, as seen in the previous section, was expected given the dominant proportion of low-risk customers. In contrast, performance on the minority (high-risk) class was substantially worse (recall: 0.26, F1-score: 0.27), with results that were similar but lower than those of the ensemble method, showing the difficulty of identifying rare high-risk cases under severe class imbalance for the Transformer model. However, overall accuracy (0.88) and the weighted metrics remained high due to the prevalence of the majority class. Both methods exhibited a similar pattern: high performance on the majority class and limited recall on the minority class.

Table 6. Classification report of the Transformer model on the test set.

Class	Precision	Recall	F1-Score	Support
False	0.94	0.94	0.94	14,250
True	0.27	0.26	0.27	1236
Accuracy			0.88	15,486
Macro avg	0.60	0.60	0.60	15,486
Weighted avg	0.88	0.88	0.88	15,486

The confusion matrix in Figure 5 shows 13,382 true negatives, 323 true positives, 868 false positives, and 913 false negatives. The performance metrics are consistent with those observed on the validation set and in the ensemble method. Using the model to exclude 8% of customers with the highest predicted risk, a real profit of 1,378,241 euros was achieved, corresponding to 38% of the maximum.

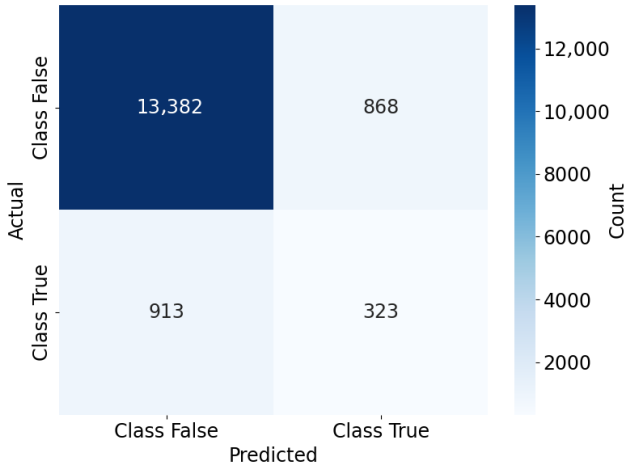

Figure 5. Confusion matrix of the Transformer model on the test set.

5.2.3. Economic Impact for the Transformer Model

Similar to the ensemble method, the economic impact analysis by customer category for the Transformer model shows the asymmetric costs associated with classification errors. As shown in Table 7, false negatives, corresponding to high-risk clients that were not detected, resulted in the largest losses per customer (−2061 euros) and a total impact of −1,869,073 euros, confirming that failing to identify these clients represents the greatest financial risk. In comparison, the ensemble method produced slightly higher per-customer costs for false negatives (−2145 euros), but both approaches faced challenges with the most costly high-risk clients. False positives, which are low-risk clients incorrectly omitted, had a much smaller average loss (147 euros) and total impact (+134,007 euros). True positives, correctly omitted high-risk clients, resulted in −1,106,242 euros in total, while true negatives generated a positive total impact of +3,247,314 euros, closely matching the profit–loss differential seen with the ensemble. Overall, these results indicate that the Transformer maintains strong performance on the majority class while still missing a non-negligible fraction of high-risk customers, which constitutes most of the economic downside.

Table 7. Economic impact analysis by customer category for the Transformer model.

Category	Count	Mean Profit (eu)	Total (eu)	Business Implication
True Negatives	13,463	243.41	+3,247,314	Correctly selected
True Positives	376	−3362.44	−1,106,242	Correctly omitted
False Negatives	860	−2060.72	−1,869,073	Missed high-risk
False Positives	787	147.42	+134,007	Foregone profit

6. Discussion

This section interprets the results from operational and actuarial perspectives. Overall, the balanced ensemble yields the most favorable trade-off between economic value, stability, and computational cost under the 8% omission constraint, whereas the Transformer provides competitive profits and improved robustness under data perturbations. Compared with the baseline methodology, both approaches shift the focus from global predictive accuracy toward portfolio value optimization under realistic underwriting limits.

6.1. Comparative Analysis Using a Real Dataset

Table 8 presents a comprehensive comparison between our balanced ensemble approach and the aforementioned baseline methodology.

Table 8. Performance comparison: balanced ensemble vs. baseline methodology.

Metric	Soriano-Gonzalez	Balanced Ensemble	Change
ROC-AUC (Test)	0.72	0.90	+25.0%
Precision (High-Risk)	0.67	0.32	−52.2%
Recall (High-Risk)	0.23	0.30	+30.4%
F1-Score (Weighted)	0.79	0.89	+12.7%
Customer Omission Rate	6.0%	7.5%	$+1.5pp$
Business Metrics			
Test Set Profit	1,232,663 euros	1,459,411 euros	+18.4%
Avg. Profit per Customer	85 euros	102 euros	+20.0%
Profit as % of Maximum	33%	40%	$+7pp$

The results show a clear improvement in the overall discrimination ability of the model when applying the balanced ensemble. The increase in ROC-AUC and the weighted

F1-score indicates that the proposed approach captures the structure of the highly imbalanced problem more effectively. This improvement is largely due to the combination of systematic undersampling and ensemble aggregation, which allows the model to learn more representative patterns from both classes.

A key difference between the two approaches appears in the treatment of the minority class. The baseline method prioritizes precision for high-risk cases, which results in fewer false positives. In contrast, the balanced ensemble achieves a noticeably higher recall, identifying more customers who truly belong to the high-risk class. In a setting where false negatives lead to much higher financial losses than false positives, this additional ability to detect costly cases is particularly important. This trade-off is expected under a strict rejection cap: increasing recall for the highest-loss customers typically requires accepting a higher false-positive rate. From a portfolio perspective, the opportunity cost of false positives is small relative to the avoided losses from correctly omitting high-severity customers.

From a business perspective, the predictive improvements translate into a clear increase in the overall profit. The balanced ensemble not only achieves a higher total profit but also increases the average profit per customer and improves the proportion of the maximum achievable profit under operational constraints. Both methods respect the business omission limit, but the proposed approach uses this capacity more efficiently. Its higher recall, together with its ability to identify economically relevant cases, helps reduce the losses associated with high-risk customers who would go unnoticed under the baseline methodology.

Table 9 presents a detailed comparison between the Transformer-based model and the methodology proposed by Soriano-Gonzalez et al. (2024). While both methods achieve comparable overall classification performance, their behaviors differ substantially across metrics relevant to business and operational objectives.

Table 9. Performance comparison: Soriano-Gonzalez methodology vs. Transformer model.

Metric	Soriano-González	Transformer	Change
ROC-AUC (Test)	0.72	0.71	−1.4%
Precision (High-Risk)	0.67	0.27	−61.1%
Recall (High-Risk)	0.23	0.26	+13.0%
F1-Score (Weighted)	0.79	0.88	+11.3%
Customer Omission Rate	6.0%	8.0%	$+2.0pp$
Business Metrics			
Test Set Profit	1,232,663 euros	1,378,241 euros	+11.8%
Avg. Profit per Customer	85 euros	96 euros	+12.9%
Profit as % of Maximum	33%	38%	$+5pp$

In terms of discrimination ability, the Transformer achieves an ROC-AUC very similar to the baseline approach, indicating that both methods separate the classes with comparable effectiveness. However, the two models behave differently when dealing with the minority class. The approach of Soriano-Gonzalez et al. (2024) maintains much higher precision, controlling false positives more effectively. In contrast, the Transformer increases recall, identifying more customers who are truly high-risk, which is especially valuable in problems where false negatives carry a high financial cost. This pattern is reflected in the improvement of the weighted F1-score, suggesting a more favorable overall balance between precision and recall.

Despite the drop in precision, the Transformer produces higher profit. The increase in total profit and average profit per customer shows that detecting more high-risk cases

compensates for the rise in false positives. Although the exclusion rate is slightly higher and fixed at 8%, this makes the approach viable in a real business setting.

6.2. Comparative Analysis with Baseline Methodology Using a Synthetic Dataset

To assess stability under controlled perturbations, multiple synthetic variants of the original dataset were generated (Synthetic000 denotes the unaltered data, and higher indices correspond to increasing noise levels). Random noise was added to the continuous numerical variables following a normal distribution with zero mean and a standard deviation proportional to the range of each variable. The resulting values were then clipped to remain within their original limits, preserving the statistical structure and inter-variable relationships while ensuring data coherence and protecting the integrity of the original information. As shown in Table 10, the index appended to each dataset name indicates the level of noise introduced into the original data, where index 0 corresponds to the unaltered dataset and higher indices represent increasing amounts of noise. Each synthetic dataset was used to train and evaluate three modeling approaches: the original methodology proposed by Soriano-Gonzalez et al. (2024), based on gradient boosting with early stopping; the Transformer-based model designed for financial risk prediction; and our balanced ensemble approach (balanced bagging with XGBoost base learners), explicitly developed to address class imbalance through systematic undersampling and ensemble aggregation.

For each model and dataset, three indicators were computed: total profit on the test set, profit percentage relative to the theoretical maximum, and average training time. The results, summarized in Table 10, reveal substantial differences among the three approaches. The Transformer model achieves high profits (ranging between 1.1 and 1.4 million euros) with an average percentage of 34%, but at the cost of very high training times (approximately 240–260 s). The balanced ensemble approach obtains comparable results in terms of profit (around 1.23 million euros and 33% on average) while significantly reducing computation time (approximately 40 to 45 s). In contrast, the original methodology yields substantially lower profits (around 0.8 million euros and 22%) but with nearly instantaneous training times. These findings highlight the trade-off between performance and computational efficiency across models and emphasize the balanced ensemble's advantage in achieving a favorable compromise between accuracy, robustness, and efficiency.

Table 10. Performance comparison across synthetic datasets for the three modeling approaches.

Dataset	Transformer			Soriano-Gonzalez et al. (2024)			Balanced Ensemble		
	Profit (eu)	%	Training Time (s)	Profit (eu)	%	Training Time (s)	Profit (eu)	%	Training Time (s)
Synthetic000	1,378,241	37	250.14	1,232,663	33	0.11	1,459,411	40	40.59
Synthetic001	1,418,905	39	235.84	940,327	26	0.11	1,214,087	33	38.09
Synthetic003	1,363,697	37	245.82	1,160,155	32	0.21	1,477,696	40	44.77
Synthetic005	1,206,787	33	237.68	1,111,400	30	0.19	1,250,344	34	46.41
Synthetic008	1,274,579	35	240.92	482,108	13	0.11	1,202,612	33	43.86
Synthetic010	1,268,976	35	239.32	938,804	26	0.21	1,109,951	30	46.51
Synthetic030	1,133,961	31	259.79	639,825	17	0.16	827,443	23	52.99
Synthetic050	1,111,445	30	243.68	406,006	11	0.09	890,799	24	37.06
Synthetic070	1,081,369	29	239.89	508,981	14	0.14	782,906	21	37.48
Synthetic090	1,112,750	30	269.36	482,103	13	0.10	857,873	23	35.19

To statistically assess whether the performance differences observed among the three modeling approaches were significant, we conducted both parametric and non-parametric analyses across the synthetic datasets. A repeated-measures ANOVA revealed a strong effect of the modeling strategy on the achieved profit ($p < 0.001$), indicating that the three approaches differ substantially. To ensure robustness to distributional assumptions, a complementary Friedman test confirmed the same pattern ($p < 0.001$, Kendall's

$W = 0.79$), demonstrating that the observed differences remain consistent even under non-parametric evaluation.

Pair-wise comparisons showed a highly consistent structure across both analysis families: the balanced ensemble model and the Transformer-based model significantly outperformed the original boosting methodology ($p < 0.01$ in all tests). No statistically significant difference was observed between the balanced ensemble and the Transformer ($p > 0.19$), indicating that both advanced approaches yield comparable profits. These results indicate that both advanced approaches deliver substantial gains in economic value relative to the baseline methodology. While the Transformer displayed slightly higher mean profits, the balanced ensemble achieved similar performance with markedly lower computational cost, making it more suitable for practical deployment.

Figure 6 illustrates the distribution of profits across synthetic datasets for the three modeling strategies. The Transformer and the balanced ensemble not only achieve higher profit values than the original approach, but also exhibit noticeably greater stability across increasing noise levels. In particular, the Transformer shows the smallest dispersion, suggesting greater robustness to perturbations and improved generalization under synthetic data variations. This reinforces the statistical results, highlighting that both modern approaches are more robust and economically effective than the original methodology.

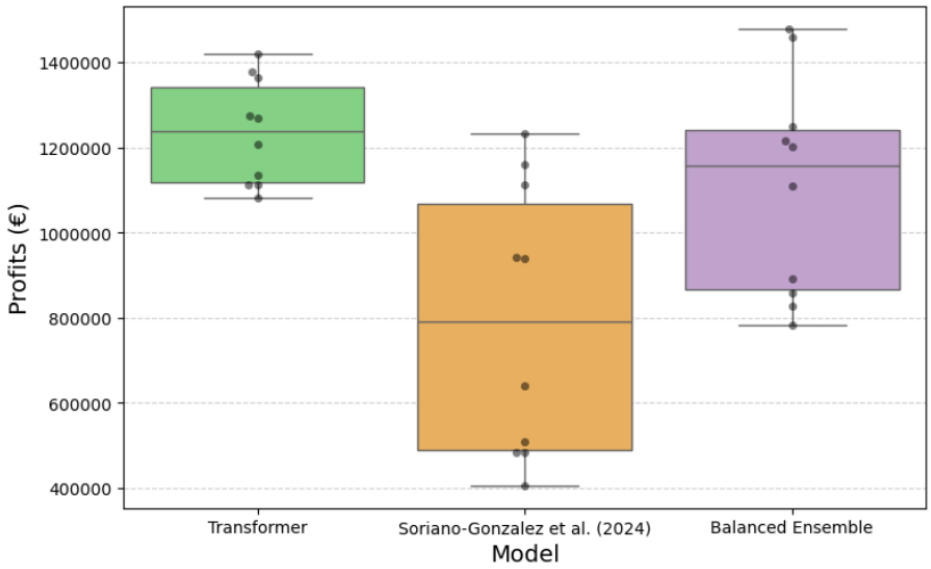

Figure 6. Distribution of profits across synthetic datasets for the three modeling strategies, including the one provided in Soriano-Gonzalez et al. (2024).

A limitation of this analysis is that the evaluation is based on a single product line and market, so performance may vary under different regulatory regimes or portfolio compositions. Nevertheless, the consistent improvements observed across synthetic perturbations suggest that the conclusions are not driven solely by dataset-specific artifacts.

7. Conclusions and Future Work

This study presented a classification framework for pseudo-new insurance customers without prior historical data, comparing two complementary approaches: a balanced XGBoost-based ensemble and a lightweight Transformer model for tabular data. Both methods operate within a profit-oriented optimization scheme that explicitly enforces the operational constraint of limiting customer omission to 8%.

From an empirical perspective, the balanced ensemble provides the most favorable trade-off between economic value, interpretability, and computational efficiency while

achieving improved detection of high-risk customers. The Transformer model delivers competitive profits and exhibits greater robustness under perturbed or synthetic data, highlighting its potential for generalization in more dynamic or data-rich environments. These findings indicate that the two approaches are not mutually exclusive but complementary: the ensemble offers a reliable and easily deployable solution in regulated contexts, whereas the Transformer represents a high-capacity alternative whose performance is expected to improve with larger datasets and increased computational resources. In addition, the ensemble structure preserves the intrinsic interpretability of tree-based models, allowing standard feature-importance analyses to identify the main underwriting drivers. This facilitates model auditing, supports regulatory compliance, and enables transparent communication of automated decisions to business stakeholders.

The main contribution of this work is the explicit alignment of the classification task with portfolio value optimization under operational constraints. By integrating asymmetric cost structures, a fixed rejection capacity, and imbalance-aware learning within a unified approach, the proposed methodology shifts the objective from global predictive accuracy toward actuarial decision-making in real underwriting scenarios. The results show that it is possible to reduce the economic impact of loss concentration without compromising portfolio growth.

A limitation of this study is that the empirical validation is based on a single product line and market. Although the synthetic experiments suggest that the conclusions are robust to controlled perturbations, model performance may vary under different behavioral patterns, regulatory frameworks, or portfolio compositions.

Future research will focus on reducing computational demands, exploring cost-sensitive adaptations to improve precision, and extending the validation to multiple markets and insurance lines. Beyond auto insurance, the approach can be transferred to life, health, and reinsurance portfolios, as well as to other domains characterized by rare and high-impact events, such as fraud detection, credit risk, and medical risk prediction.

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

Funding: This work has been partially supported by the Spanish Ministry of Science, Innovation and Universities/AEI (PID2022-138860NB-I00, DIN2024-013395, AIA2025-163553-C44) and the Generalitat Valenciana (2024 CIAICO 117).

Data Availability Statement: The benchmark dataset presented in this paper is available at https://doi.org/10.6084/m9.figshare.31894816, which provides direct and unrestricted access to the data.

Conflicts of Interest: Author Antoni Guerrero was employed by the company Baobab AI. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict 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 Oversampling Technique

References

Anderson, Theodore W., and Donald A. Darling. 1952. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics* 23: 193–212. [CrossRef]

Baran, Sebastian, and Przemysław Rola. 2022. Prediction of motor insurance claims occurrence as an imbalanced machine learning problem. *arXiv* arXiv:2204.06109. [CrossRef]

Black, Julia, and Andrew Douglas Murray. 2019. Regulating ai and machine learning: Setting the regulatory agenda. *European Journal of Law and Technology* 10: 1–21

Brati, Esmeralda, Alma Braimllari, and Ardit Gjeçi. 2025. Machine learning applications for predicting high-cost claims using insurance data. *Data* 10: 90. [CrossRef]

Bruun, Simone Borg, Christina Lioma, and Maria Maistro. 2024. Recommending target actions outside sessions in the data-poor insurance domain. *ACM Transactions on Recommender Systems* 3: 1–24. [CrossRef]

Cambria, Erik, and Bebo White. 2014. Jumping nlp curves: A review of natural language processing research. *IEEE Computational Intelligence Magazine* 9: 48–57. [CrossRef]

Chawla, Nitesh V., Aleksandar Lazarevic, Lawrence O. Hall, and Kevin W. Bowyer. 2003. Smoteboost: Improving prediction of the minority class in boosting. In *European Conference on Principles of Data Mining and Knowledge Discovery*. Berlin/Heidelberg: Springer, pp. 107–19.

Chen, Tianqi, and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: Association for Computing Machinery, pp. 785–94.

Clauset, Aaron, Cosma Rohilla Shalizi, and Mark E. J. Newman. 2009. Power-law distributions in empirical data. *SIAM Review* 51: 661–703. [CrossRef]

De Meulemeester, Hannes, and Bart De Moor. 2020. Unsupervised embeddings for categorical variables. In *2020 International Joint Conference on Neural Networks (IJCNN)*. Piscataway: IEEE, pp. 1–8.

Dietterich, Thomas G. 2000. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*. Berlin/Heidelberg: Springer, pp. 1–15.

Doshi-Velez, Finale, and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv* arXiv:1702.08608. [CrossRef]

Elbhrawy, Ahmed Shawky, Mohamed AbdelFattah Belal, and Mohamed Sameh Hassanein. 2024. Ces: Cost estimation system for enhancing the processing of car insurance claims. *Journal of Computing and Communication* 3: 55–69. [CrossRef]

Finger, Dina, Hansjoerg Albrecher, and Lutz Wilhelmy. 2024. On the cost of risk misspecification in insurance pricing. *Japanese Journal of Statistics and Data Science* 7: 1111–53. [CrossRef]

Fitriani, Maulida Ayu, and Dany Candra Febrianto. 2021. Data mining for potential customer segmentation in the marketing bank dataset. *JUITA: Jurnal Informatika* 9: 25–32. [CrossRef]

Galar, Mikel, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. 2012. 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: 463–84. [CrossRef]

Ghorbani, Elnaz, and Raquel Soriano-Gonzalez. 2024. Enhancing predictive models in insurance: A feature selection analysis. In *Decision Science Alliance International Summer Conference*. Berlin/Heidelberg: Springer, pp. 256–67.

Giudici, Paolo, and Laura Parisi. 2018. Corisk: Credit risk contagion with correlation network models. *Risks* 6: 95. [CrossRef]

Gorishniy, Yury, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. 2021. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems*. Cambridge: The MIT Press, vol. 34, pp. 18932–43.

Gupta, Vibhuti, Julian Broughton, Ange Rukundo, and Lubna Pinky. 2025. Learning unbiased risk prediction based algorithms in healthcare: A case study with primary care patients. *Informatics in Medicine Unlocked* 54: 101627. [CrossRef]

Hanafy, Mohamed, and Ruixing Ming. 2021. Machine learning approaches for auto insurance big data. *Risks* 9: 42. [CrossRef]

Henckaerts, Roel, Marie-Pier Côté, Katrien Antonio, and Roel Verbelen. 2021. Boosting insights in insurance tariff plans with tree-based machine learning methods. *North American Actuarial Journal* 25: 255–85. [CrossRef]

Hu, Changyue, Zhiyu Quan, and Wing Fung Chong. 2022. Imbalanced learning for insurance using modified loss functions in tree-based models. *Insurance: Mathematics and Economics* 106: 13–32. [CrossRef]

Hutagaol, B. Junedi, and Tuga Mauritsius. 2020. Risk level prediction of life insurance applicant using machine learning. *International Journal of Advanced Trends in Computer Science and Engineering* 9: 2213–20. [CrossRef]

Jarque, Carlos M., and Anil K. Bera. 1987. A test for normality of observations and regression residuals. *International Statistical Review* 55: 163–72. [CrossRef]

Khamesian, Farzan, Maryam Esna-Ashari, Eric Dei Ofosu-Hene, and Farbod Khanizadeh. 2022. Risk classification of imbalanced data for car insurance companies: Machine learning approaches. *International Journal of Mathematical Modelling & Computations* 12: 153–62.

Kinga, Diederik, and Jimmy Ba Adam. 2015. A method for stochastic optimization. Paper presented at the International Conference on Learning Representations (ICLR), San Diego, CA, USA, May 7–9, vol. 5.

Kolambe, Sapana, and Parminder Kaur. 2023. Survey on insurance claim analysis using natural language processing and machine learning. *International Journal on Recent and Innovation Trends in Computing and Communication* 11: 30–38. [CrossRef]

Krupova, Marketa, Nabil Rachdi, and Quentin Guibert. 2025. Explainable boosting machine for predicting claim severity and frequency in car insurance. *arXiv* arXiv:2503.21321. [CrossRef]

Le, Thi-Thu-Huong, Aji Teguh Prihatno, Yustus Eko Oktian, Hyoeun Kang, and Howon Kim. 2023. Exploring local explanation of practical industrial ai applications: A systematic literature review. *Applied Sciences* 13: 5809. [CrossRef]

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521: 436–44. [CrossRef]

Leo, Martin, Suneel Sharma, and Koilakuntla Maddulety. 2019. Machine learning in banking risk management: A literature review. *Risks* 7: 29. [CrossRef]

Li, Yu, Yi Zhang, Lu Gan, Gengwei Hong, Zimu Zhou, and Qiang Li. 2021. Revman: Revenue-aware multi-task online insurance recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Washington: AAAI Press, vol. 35, pp. 303–10.

Lim, Sehyun, Taeyeon Oh, and Guy Ngayo. 2023. Analyzing factors affecting risk aversion: Case of life insurance data in korea. *Heliyon* 9: e20697. [CrossRef]

Liu, Xu-Ying, Jianxin Wu, and Zhi-Hua Zhou. 2008. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39: 539–50. [CrossRef]

Loisel, Stéphane, Pierrick Piette, and Cheng-Hsien Jason Tsai. 2021. Applying economic measures to lapse risk management with machine learning approaches. *ASTIN Bulletin: The Journal of the IAA* 51: 839–71. [CrossRef]

Lundberg, Scott M., and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*. Cambridge: The MIT Press, vol. 30.

Marhavilas, Panagiotis K., and Dimitrios E. Koulouriotis. 2021. Risk-acceptance criteria in occupational health and safety risk-assessment—The state-of-the-art through a systematic literature review. *Safety* 7: 77. [CrossRef]

Masello, Leandro, German Castignani, Barry Sheehan, Montserrat Guillen, and Finbarr Murphy. 2023. Using contextual data to predict risky driving events: A novel methodology from explainable artificial intelligence. *Accident Analysis & Prevention* 184: 106997. [CrossRef]

Mienye, Ibomoiye Domor, and Theo G. Swart. 2024. A hybrid deep learning approach with generative adversarial network for credit card fraud detection. *Technologies* 12: 186. [CrossRef]

Ngwenduna, Kwanda Sydwell, and Rendani Mbuvha. 2021. Alleviating class imbalance in actuarial applications using cost-sensitive learning. *Risks* 9: 49. [CrossRef]

Olasehinde, Olayemi, Boniface Kayode Alese, and Ojonukpe Eqwuche. 2025. Privacy-preserving artificial intelligence: Principles, methods, applications, and challenges. *Journal of Applied Artificial Intelligence* 6: 60–70. [CrossRef]

Orji, Ugochukwu, and Elochukwu Ukwandu. 2024. Machine learning for an explainable cost prediction of medical insurance. *Machine Learning with Applications* 15: 100516. [CrossRef]

Pes, Barbara. 2021. Learning from high-dimensional and class-imbalanced datasets using random forests. *Information* 12: 286. [CrossRef]

Sadreddini, Zhaleh, Ilknur Donmez, and Halim Yanikomeroglu. 2021. Cancel-for-any-reason insurance recommendation using customer transaction-based clustering. *IEEE Access* 9: 39363–74. [CrossRef]

Sahai, Rahul, Ali Al-Ataby, Sulaf Assi, Manoj Jayabalan, Panagiotis Liatsis, Chong Kim Loy, Abdullah Al-Hamid, Sahar Al-Sudani, Maitham Alamran, and Hoshang Kolivand. 2022. Insurance risk prediction using machine learning. In *The International Conference on Data Science and Emerging Technologies*. Berlin/Heidelberg: Springer, pp. 419–33.

Sari, Puspita Kencana, and Adelia Purwadinata. 2019. Analysis characteristics of car sales in e-commerce data using clustering model. *Journal of Data Science and Its Applications* 2: 19–28. [CrossRef]

Seiffert, Chris, Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. 2009. Rusboost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 40: 185–97. [CrossRef]

Seyam, Eslam Abdelhakim. 2025. Predicting high-cost healthcare utilization using machine learning: A multi-service risk stratification analysis in eu-based private group health insurance. *Risks* 13: 133. [CrossRef]

Shah, Chiranjibi, Qian Du, and Yan Xu. 2022. Enhanced tabnet: Attentive interpretable tabular learning for hyperspectral image classification. *Remote Sensing* 14: 716. [CrossRef]

Simester, Duncan, Artem Timoshenko, and Spyros I. Zoumpoulis. 2020. Targeting prospective customers: Robustness of machine-learning methods to typical data challenges. *Management Science* 66: 2495–522. [CrossRef]

Singla, Jimmy, Ali Kashif Bashir, Yunyoung Nam, Najam UI Hasan, and Usman Tariq. 2021. Handling class imbalance in online transaction fraud detection. *Computers, Materials and Continua* 70: 2861–77.

Soriano-Gonzalez, Raquel, Veronika Tsertsvadze, Celia Osorio, Noelia Fuster, Angel A. Juan, and Elena Perez-Bernabeu. 2024. Balancing risk and profit: Predicting the performance of potential new customers in the insurance industry. *Information* 15: 546. [CrossRef]

Tian, Xiaoguang, Jun Todorovic, and Zelimir Todorovic. 2023. A machine-learning-based business analytical system for insurance customer relationship management and cross-selling. *Journal of Applied Business & Economics* 25: 273–89.

Tiwari, Anoop Kumar, Rajat Saini, Abhigyan Nath, Phool Singh, and Mohd Asif Shah. 2024. Hybrid similarity relation based mutual information for feature selection in intuitionistic fuzzy rough framework and its applications. *Scientific Reports* 14: 5958. [CrossRef]

Uddin, Moin, Mohd Faizan Ansari, Mohd Adil, Ripon K. Chakrabortty, and Michael J. Ryan. 2023. Modeling vehicle insurance adoption by automobile owners: A hybrid random forest classifier approach. *Processes* 11: 629. [CrossRef]

Voigt, Paul, and Axel Von dem Bussche. 2017. The EU general data protection regulation (gdpr). *A Practical Guide*, 1st ed. Cham: Springer International Publishing, vol. 10, pp. 10–5555.. [CrossRef]

Young, Tom, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. Recent trends in deep learning based natural language processing. *IEEE Computational IntelligenCe Magazine* 13: 55–75. [CrossRef]

Zhou, Zhi-Hua. 2012. *Ensemble Methods: Foundations and Algorithms*. Boca Raton: CRC Press.

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.