Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/200413

Modeling Extreme Risks in Non-Life Insurance, from Threshold Estimation to AI Premiums

Yassine Kouach[1,*] *and Abderrahim EL Attar*[2]

[1]National School of Business and Management of El Jadida (ENCG El Jadida), University Chouaib Doukkali, El Jadida, Morocco
[2]Faculty of Law, Economic and Social Sciences Ain Sebaa, University Hassan II, Casablanca, Morocco

Received: 12 May 2026, Revised: 2 Jun. 2026, Accepted: 22 Jun. 2026
Published online: 1 Jul. 2026

Abstract: Pricing extreme risks is difficult given the rarity of these events, the small volume of severe observations, and the peculiarity of their statistical distribution. Thus, traditional methods such as generalized linear models are unable to model the frequency and severity of extreme claims more accurately. On the other hand, most of the work put forward in the literature has required the formulation of assumptions on the distribution of extreme data, which depend on the threshold of severe claims used. In this regard, this study proposes a new approach to pricing extreme disasters based on artificial intelligence instead of the models traditionally used in this type of modeling. The proposed approach makes it possible to exploit artificial intelligence throughout the pricing process for extreme risks, from identifying the threshold to determining the pure premium. To achieve this, a k-means algorithm is used to determine the threshold for extreme losses. The occurrence of these losses is predicted by using machine learning methods adapted to classification problems. An actuarial model of pure premium insurance is developed to accommodate extreme losses, while the frequency and average cost of major losses are estimated using machine learning algorithms. The results of this study demonstrate the effectiveness of the proposed approach, based on artificial intelligence algorithms, in offering fair prices tailored to the high level of risk involved. This approach allows insurance companies to optimize the pricing of high-risk insurance contracts.

Keywords: Artificial Intelligence, Extreme risks, Extreme value theory, Machine Learning, Non-Life Insurance, Pure Premium Prediction, Threshold estimation

1 Introduction

Insurance companies place great importance on modeling extreme risks such as fires and natural disasters due to the colossal compensation costs in the event of claims. Therefore, insurance companies are seeking to better price these products at the right price to be able to cope with the losses inherent to these risks, to achieve a profit margin, and to ensure the company's solvency.

However, pricing extreme risks is difficult given the rarity of events, the small volume of data and the peculiarity of their statistical distribution. Thus, traditional methods such as generalized linear models are unable to model the frequency and severity of extreme losses more accurately.

According to the theory of extreme values, a loss of an amount above a threshold is qualified as an extreme claim and if the amount is lower than this threshold by an attritional claim. However, the problem that persists is the choice of the extreme claim's threshold, and therefore the type of extreme claims modeling. In this perspective, a fairly large number of research works have been initiated.

Among these works, [1] propose two approaches to model asymmetric insurance claims data containing large extreme values. The first approach is to use the generalized lambda distribution, while the second approach is based on the use of the semi-parametric transformed kernel density. Their results demonstrate the advantage of these approaches over the generalized Pareto distribution [1].

[2] use a parametric approach to model the severity of fire losses in Norway. For this, [2] consider a set of probability distributions, including the generalized Pareto distribution as well as other models proposed more recently in actuarial literature, such as the log-normal-Pareto and Weibull-Pareto distributions.

[3] develop a severity threshold model that divides the claim size distribution into two zones, a lower zone, where the claim size is below a given threshold, and an

* Corresponding author e-mail: kouach.yassine@ucd.ac.ma

© 2026 NSP
Natural Sciences Publishing Cor.

upper zone, where the claim size exceeds this threshold [3]. [3] model extreme claims above the threshold according to the Peaks-Over-Threshold (POT) methodology of extreme value theory, using the generalized Pareto distribution for the excess distribution [3]. Claims below the threshold are modeled by a generalized linear model based on the truncated gamma distribution [3]. [3] demonstrate the advantage of the severity threshold model over the commonly used generalized linear model based on the gamma distribution.

[4] examine various composite models with commonly used global distributions combined with a two-parameter Pareto distribution above a certain threshold [4]. [4] explore the effectiveness of several threshold selection methods in reserve estimation, as well as the effect of choosing the global distribution, depending on sample sizes and varying tail properties [4]. The study by [4] reveals that, when data are sufficient, the empirical rule generally performs best in terms of reserve estimation quality [4].

[5] propose a method for analyzing cyber claims using regression trees to identify criteria for classifying and evaluating claims [5]. They focus on extreme claims [5], combining generalized Pareto modeling from extreme value theory with a regression tree approach [5]. Coupled with a frequency assessment, their procedure allows for the calculation of central scenarios and extreme loss quantiles for a cyber portfolio [5].

[6] focus on one of the methods of extreme value theory, namely "Peaks-Over-Threshold" (POT), where the thresholds of a generalized Pareto distribution (GPD) are determined from a Hill diagram [6]. [6] propose a new approach to select k via the Hill estimator obtained using a quantile function of type 8 [6]. [6] use the obtained threshold to identify the Value at Risk (VaR) and the expected deficit for an event.

[7] explore the estimation of the Expectile conditional on extreme tails using quantile regression and Expectile regression models [7]. [7] propose three methods to perform extrapolation based on a second-order condition in a framework of so-called conditionally heteroscedastic and unconditionally homoscedastic extremes [7].

[8] develop a methodology to automatically classify claims into severe or not, based on the information contained in text reports [8]. [8] propose different rebalancing algorithms to solve the problem of imbalanced data, where few observations are associated with the extreme claim label. In addition, they use different methodologies to process text data from reports [8].

[9] Seek to apply extreme value theory to actuarial science using an innovative parameter estimator, allowing the calculation of pure reinsurance premiums and the choice of retention limit for insurance companies [9]. [9] conclude that extreme value theory is a powerful tool for insurance, better capturing the behavior of extreme claims compared to traditional models [9].

[10] seek to improve the ability of deep learning to model extreme events for time series prediction. [10] find that the weaknesses of deep learning methods stem from the conventional form of quadratic loss. To address this issue, [10] draw inspiration from extreme value theory and develop a new loss function, called Extreme Value Loss (EVL), to detect the future occurrence of extreme events. Furthermore, [10] propose the use of Memory Networks to memorize extreme events from historical data. [10] develop an end-to-end framework for time series prediction with extreme events, by integrating Extreme Value Loss with a suitable Memory Network module.

Most of the work is based on Extreme Value Theory (EVT), that is the theoretical framework for models analyzing extreme values. Some models, such as Log-Normal, Weibull, and Pareto [1], fail to model extreme observations of atypical nature. In contrast, several models proposed in the literature use the Generalized Pareto Distribution (GPD) to model extreme claims. However, these approaches require the formulation of assumptions on data distribution, which depend on a threshold that separates extreme claims from attritional claims.

In this perspective, this work presents a new approach to pricing extreme risks based on exploiting the capacity of artificial intelligence tools as a substitute for traditional methods throughout the pricing process. Indeed, this approach makes it possible to determine the threshold of extreme claims using unsupervised machine learning methods, in this case the K-means algorithm. In addition, we model the occurrence of extreme claims using machine learning methods. We also propose actuarial modeling tailored to the pure premium for extreme losses, while exploiting the performance of artificial intelligence to determine the average cost and frequency of severe claims.

To put our extreme loss pricing approach into practice, we determine the pure premiums for an insurance product covering fire risk. The results demonstrate the advantages of pricing extreme claims using artificial intelligence over traditional methods, allowing us to offer fair-value premiums based on the degree of risk assumed. The implications of this study contribute to actuarial science, particularly regarding pricing issues within the framework of extreme value theory.

This article is developed as follows. In the second section, we present our methodology adopted to propose our approach to pricing extreme claims. In this methodology section, we present the theoretical framework of extreme claims modeling. We also proceed to the theoretical development of our approach that consists of defining the actuarial model of the technical premium of extreme claims, the severity threshold model, the extreme claims prediction model and the pure premium prediction model. We present the results of this

© 2026 NSP
Natural Sciences Publishing Cor.

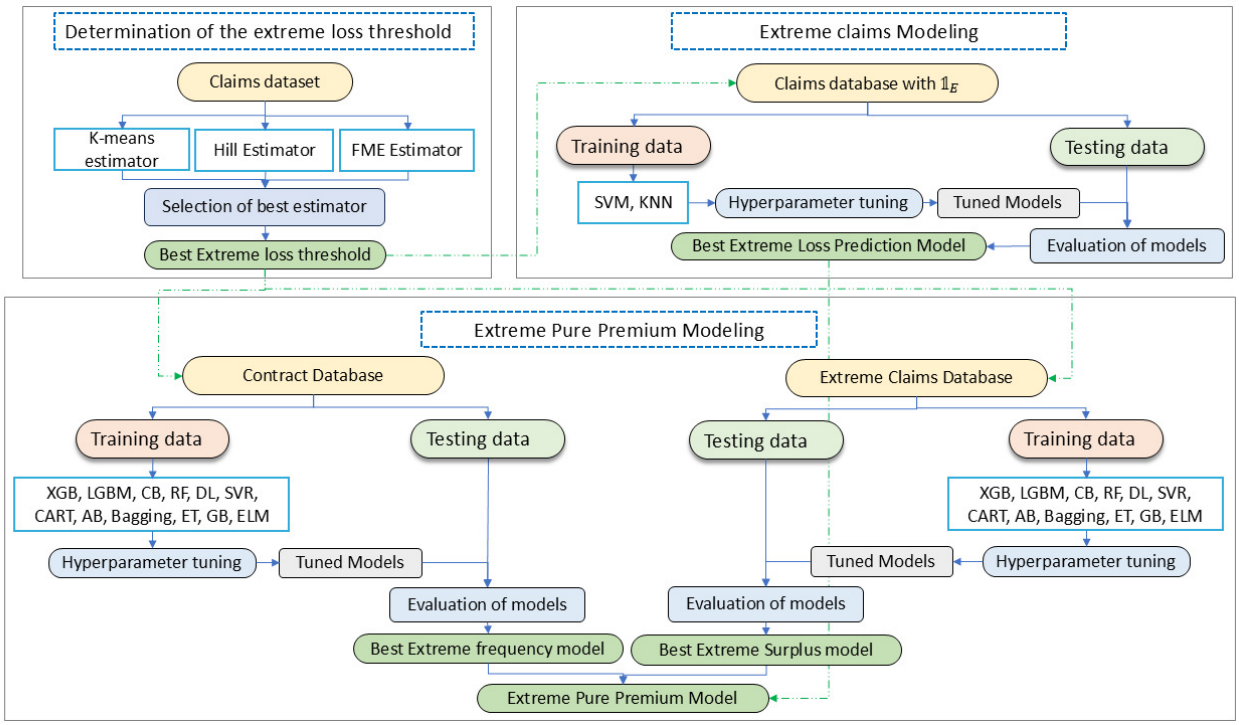

Fig. 1: Extreme Risk Pricing Diagram (by us)

study in the penultimate section. And finally, we close this work with a conclusion in the last section.

2 Methodology

2.1 Proposed approach for pricing extreme claims

In order to develop our extreme claims pricing model, we establish three essential phases as illustrated in Fig. 1, namely the phase of determining the extreme claims threshold, the phase of modeling extreme claims occurrence and the phase of modeling the pure premium associated with the extreme claim.

The first phase consists of determining the extreme loss threshold t. To do this, we use two methods commonly used in extreme value theory, namely the Hill estimator and the mean excess function estimator.

In this article, we propose to use an alternative method derived from artificial intelligence, in this case an unsupervised statistical learning method, namely K-means, to determine the extreme loss threshold.

This threshold will be used in the second phase, which consists of modeling extreme losses. We construct a new claims database with a new variable indicating that the loss is normal if the loss amount is below the estimated threshold, or that the loss is extreme when the loss amount is above the estimated threshold.

We then develop an extreme loss prediction model, which consists of modeling the occurrence of extreme losses in order to estimate the indicator variable using the new claims database.

The nature of fire insurance claims allows us to observe that the insured can suffer either a normal loss or an extreme loss. The insured cannot suffer more than one extreme loss per year of coverage. Indeed, it is unlikely that there will be two serious losses in a single year in fire insurance due to the time taken to assess and restore the property to its initial condition following an extreme fire. Based on this observation, our extreme loss prediction model is a binary classification. This model then consists of predicting whether the insured may suffer an extreme loss $Y = \mathbb{1}_E = 1$, or not $Y = \mathbb{1}_E = 0$.

To do this, we use five machine learning algorithms, namely the Support Vector Machine algorithm, the K Nearest Neighbors algorithm, the Isolation Forest algorithm, the One Class SVM algorithm, and the Auto-encoder algorithm. In order to choose the most appropriate model for predicting extreme losses, we evaluate these models based on performance measures adapted to a binary classification (Table 1). The model elected as efficient will be used as a model for predicting the occurrence of extreme losses.

We then divide the database into a database with normal claims and a database with severe claims. The first database will be used to model the average cost and the frequency of attritional claims, and the second database

© 2026 NSP
Natural Sciences Publishing Cor.

Table 1: Performance measures of a binary predictive model

Measure	Description
$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$	Accuracy measures the proportion of correctly classified observations among all observations [11].
$Precision = \frac{TP}{TP+FP}$	Precision measures the proportion of correctly predicted positive observations among all predicted positive observations [11].
$Recall = \frac{TP}{TP+FN}$	Recall measures the proportion of actual positive observations that are correctly identified by the model [11].
$Specificity = \frac{TN}{TN+FP}$	Specificity measures the proportion of correctly identified negative observations among all actual negative observations [11].
$F1\text{-}Score = 2\frac{Precision \times Recall}{Precision+Recall}$	The F1-score combines Precision and Recall into a single metric, providing a balanced evaluation of the model performance, especially for imbalanced datasets.

will be used to model the pure extreme premium in the third phase.

The pure extreme premium corresponds to the increase that the policyholder must pay if they are at risk of suffering an extreme claim by determining the frequency of an extreme claim and the average amount of the difference between the extreme claim amount and the extreme claim threshold.

With this in mind, we add a new variable to the second database of severe claims, which indicates the difference between the extreme claim amount and the extreme claim threshold. This variable is used to model the extreme excess. Furthermore, it is necessary to merge the second database, including the excess variable, with the contract database, adding another variable for the frequency of extreme claims to develop the extreme frequency model.

We use a basket of machine learning algorithms to develop the extreme frequency model and the extreme surplus model. These algorithms are the XG Boost algorithm, the Light GBM algorithm, the Cat Boost algorithm, the Random Forest algorithm, the Deep Learning algorithm, the Support Vector Machine Regressor SVR algorithm, the Classification And Regression Tree CART algorithm, the AdaBoost algorithm, the Bagging algorithm, the Extra Trees algorithm, the Gradient Boosting algorithm, and the ELM Extreme Learning Machine algorithm.

In order to measure the performance of these algorithms, we use Root Mean Square Error (RMSE) the most used in the context of regression.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

In the different phases of our approach, we divide each database used into two subsets, a training set and an evaluation set with respective proportions of 70% and 30%. Thus, we use the cross-validation technique which allows us to search for the optimal values of the hyperparameters of the algorithms.

2.2 Description of the pure premium model for extreme claims

The famous approach to *P&C* Property and Casualty (Fire, Accident and Miscellaneous Risks) insurance pricing, in this case the "Frequency * Average Cost" method, consists first of all in determining the technical premium or in other words the pure premium which corresponds to the cost of the risk.

The pure premium is only the expectation of the total claims burden.

$$pure \quad premium = E[S] \tag{2}$$

where S is a random variable representing the total claims load [12].

In order to determine the law of the random variable S, there are two models in the actuarial literature, namely the individual model and the collective model [12]. According to the collective model consists of analyzing all claims in a combined manner. For this, the total load of claims is then written:

$$S_{coll} = \sum_{i=1}^{n} X_i \tag{3}$$

where: N is a random variable representing the number of declared claims, and X_i is a random variable designating the cost of claim i. Using the collective model and assuming that the number of claims N is independent of the cost of claims and that the random variables of the costs of claims have the same distribution, we can write the expectation of the total load of claims according to the famous model in *P&C* insurance pricing, in this case the "Frequency $*$ Average Cost" method [12]:

$$E[S] = E[N] * E[X] \tag{4}$$

This approach consists in modeling the pure premium by estimating separately the average number of claims $E[N]$ and the average cost of claims $E[X]$ [12,13] in order to multiply them afterwards.

$$\begin{aligned}\text{Pure Premium} &= \text{Average Number of Claims} \\ &\times \text{Average Cost of Claims}\end{aligned} \tag{5}$$

© 2026 NSP
Natural Sciences Publishing Cor.

However, this modeling only allows pricing for attritional claims. It does not take into account the pricing of extreme claims. To do this, we propose an approach that makes it possible to integrate the pricing of extreme claims into the calculation of the pure premium.

To qualify a disaster as extreme, it must exceed a threshold. Therefore, by exceeding this threshold, the insured is no longer within the framework of normal claims. To this end, this category of insured must be penalized by an additional premium that compensates for the higher risk incurred by the insurer.

In other words, if an extreme loss is predicted for the insured, the insured will pay a pure premium composed of two premiums. A pure premium for attritional losses that are below the chosen threshold, and an extreme pure premium for the excess, which corresponds to the difference between extreme losses and the chosen threshold.

From this perspective, we model the pure premium in the following form, keeping the same assumptions made in the frequency-cost model:

$$E[S] = E[N] * E[X] + 1_E * E[N_E] * E[X_E - t] \quad (6)$$

Where $E[N]$ is the average number of normal claims, $E[X]$ is the average cost of normal claims, 1_E is an indicator variable for the occurrence of the extreme claim,($1_E = 0$ if the loss is normal and $1_E = 1$ if the loss is extreme), $E[N_E]$ is the average number of extreme claims and $E[X_E - t]$ is the average excess amount resulting from the difference between the amount of the extreme loss X_E and the extreme claim threshold t.

$$\begin{aligned}\text{Pure Premium} = \text{Attritionnel pure premium} \quad + \\ 1_E \times \text{Extreme pure premium}\end{aligned} \quad (7)$$

If the policyholder is not at risk of suffering a serious loss, they will only pay the pure attritional premium. However, if the policyholder is potentially at risk of suffering an extreme loss, they will pay the pure attritional premium plus a surplus in the form of a pure extreme premium, which represents the average amount of the excess of extreme losses over the serious loss threshold.

The calculation of the pure premium, taking extreme losses into account, involves performing four models: modeling the frequency of attritional losses, modeling the average cost of attritional losses, modeling the frequency of extreme losses, and modeling the average cost of the excess of extreme losses.

2.3 Theoretical framework for modeling extreme losses

Extreme value theory (EVT) focuses on the study of the tail of a claims distribution, that represents the highest claim amounts, and measuring the probability of occurrence in high quantiles, unlike classical statistical theory, that focuses mainly on studying the behavior of values around the mean or median of a distribution [14, 15].

Indeed, extreme value theory (EVT) makes it possible to isolate large (extreme) values that are random and rare events but which have a potentially significant impact [16], and to model them separately from the rest of the distribution, in order to improve their prediction as well as that of ordinary values [17].

There are two main extreme value theory approaches used to model extreme values, namely the generalized extreme value (GEV) approach based on the maxima (minima) block [18] and the generalized Pareto distribution (GP) method based on the POT Peaks-Over-Threshold approach [19,20].

The block-based approach models the maximum or minimum of a very large sample, while the peak-over-threshold (POT) approach studies the distribution of excess losses above a threshold.

2.3.1 Maxima (minima) block approach

The maximum limit law approach was developed through the Fisher-Tippett extreme value theorem [21], the work of Gumbel [22], and the work of Gnedenko [23] that provided the final version of the extreme value theorem. Let be $X_1, ..., X_n$ a sequence of i.i.d. random variables with distribution function F, with $M_n = max(X_i)_{1 \le i \le n} = max(X_1, ..., X_n)$ is the maximum. We denote F the distribution function of the variable X. Then [24]

$$\begin{aligned} F_{M_n} &= Pr(M_n \le x) \\ &= Pr(X_1 \le x, ..., X_n \le x) \\ &= Pr(X_1 \le x) * ... * Pr(X_n \le x) \\ &= P(\bigcap_{i=1}^{n} X_i \le x) \\ &= P(X_1 \le x)^n \\ &= [F(x)]^n \end{aligned} \quad (8)$$

We are interested in an approximation of the asymptotic distribution of the maximum.

$$\lim_{n\to\infty} F_{M_n}(x) = \lim_{n\to\infty} [F(x)]^n = \begin{cases} 0, & \text{if } x < x_F, \\ 1, & \text{otherwise,} \end{cases} \quad (9)$$

where $x_F = \sup\{x \in \mathbb{R} : F(x) < 1\}$ is the extreme point of the distribution function F [25]. Then, $M_n \to x_F$ in probability when $n \to \infty$.

Fischer–Tippett Theorem [21]

© 2026 NSP
Natural Sciences Publishing Cor.

Assume n i.i.d. random variables $X_1,\ldots,X_n$ with distribution function F. If there exist two real sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$ with $b_n \geq 0$, and a non-degenerate distribution function G [26] such that

$$\frac{M_n - a_n}{b_n} \longrightarrow G, \tag{10}$$

when n tends to infinity, then G is of the same type as one of the following three laws [23,24]:

–Fréchet's law

$$\Phi_\alpha(x) = \begin{cases} \exp(-x^{-\alpha}), & x > 0, \\ 0, & \text{otherwise}, \end{cases} \qquad \alpha > 0. \tag{11}$$

[27]

The Fréchet distribution was introduced by the French mathematician Maurice René Fréchet [28] as a law of extreme values.

–Weibull's law

$$\Psi_\alpha(x) = \begin{cases} \exp(-(-x)^{-\alpha}), & x < 0, \\ 1, & \text{otherwise}, \end{cases} \qquad \alpha < 0. \tag{12}$$

The Weibull distribution was introduced by the Swedish mathematician Waloddi Weibull [29].

–Gumbel's law

$$\gamma(x) = \exp(-\exp(-x)) \tag{13}$$

The Gumbel distribution was introduced by the American mathematician Emil Julius Gumbel as an extreme value distribution [22].

Generalized extreme value (GEV) distribution

The three distributions, namely Fréchet, Gumbel, and Weibull, are special cases of the most general form of the generalized extreme value distribution. [30] proposed the generalized extreme value distribution (GEV), which allows the three distributions to be combined into a single distribution [31], [32]. The distribution function of a GEV is generally expressed as [33]:

$$G_{\mu,\sigma,\xi}(x) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \tag{14}$$

with μ the position parameter, σ the dispersion parameter, ξ the shape parameter (or the tail index, which depends only on the law of X).

The possible limiting distributions are parameterized by the extreme value index ξ, that makes it possible to divide the limiting distributions into three domains of attraction, Fréchet, Weibull and Gumbel [14]: - If $\xi = \alpha^{-1} > 0$, the distribution is **Fréchet** (heavy tail). - If $\xi = \alpha^{-1} < 0$, the distribution is **Weibull** (light tail). - If $\xi = 0$, the distribution is **Gumbel** (intermediate tail).

To estimate the GEV, we assume that the limiting distribution is approximated by:

$$P\left(\frac{M_n - a_n}{b_n} \leq x\right) = [F(a_n x + b_n)]^n \longrightarrow GEV(\mu,\sigma,\xi) \tag{15}$$

Thus, to model the maximum limit distribution, it is sufficient to estimate it directly using a GEV distribution.

The GEV parameters are estimated using the "Block Maxima" method. To do this, it is necessary to ensure that the blocks are sufficiently large.

2.3.2 The Peaks-Over-Threshold (POT) approach

The threshold exceedance approach, proposed by [34], consists of choosing a threshold high enough to determine the excess values that exceed this threshold, and modeling these excesses by an asymptotic distribution given by Pickands' theorem [34].

Law of Excesses

The threshold exceedance method is used to model extreme values by approximating the law of excesses exceeding a threshold u.

Let $X-u$ an excess be above a threshold u. The excess value represents the difference between the value exceeding the threshold u and the threshold u [25]. This threshold u must be sufficiently large and less than the maximum value of the actual random variable X [25,35].

The law of excesses is approximated using the conditional law of the strictly positive random variable $X-u$ (with $X > u$). In fact, we seek to determine a conditional distribution F_u of the random variable $X-u$ from the distribution F_X [36]. The conditional distribution function F_u of excesses beyond the threshold u is defined by:

$$\begin{aligned} F_u(y) &= P(X-u \leq y \mid X > u) \\ &= \begin{cases} \dfrac{F(u+y)-F(u)}{1-F(u)}, & \text{si } y \geq 0, \\ 0, & \text{si } y < 0 \end{cases} \end{aligned} \tag{16}$$

Generalized Pareto Law

Pickands' theorem [34] allows us to determine the form of the limiting distribution of excesses that can converge to a generalized Pareto distribution (GPD) under

© 2026 NSP
Natural Sciences Publishing Cor.

certain conditions. According to the Pickands–Balkema–De Haan theorem [27,34,37], a distribution function F belongs to the domain of maximum attraction G_ξ if and only if there exists a positive function σ and a real number ξ such that [38]:

$$\lim_{u\to x_F} \sup_{0\le y\le x_F-u} \left| F_u(y) - F^{GPD}_{\xi;\sigma(u)}(y) \right| = 0 \tag{17}$$

where $F_u(y)$ is the conditional excess distribution function for u high, x_F is the extreme point of F [38], $x_f = \sup\{x \in \mathbb{R} : F(x) < 1\}$ and $F^{GPD}_{\xi,\sigma(u)}(y)$ is the Generalized Pareto function.

The Generalized Pareto function is defined by:

$$F^{GPD}_{\xi,\sigma(u)}(y) = \begin{cases} 1 - \left[1 + \xi\left(\frac{y}{\sigma(u)}\right)\right]^{-1/\xi}, & \text{si } \xi \neq 0, \\ 1 - e^{-y/\sigma(u)}, & \text{si } \xi = 0, \end{cases} \tag{18}$$

where $y \ge 0$ if $\xi \ge 0$; and $0 \le y \le -\sigma(u)/\xi$ if $\xi < 0$.

The generalized Pareto distribution is denoted $GPD(\xi, \sigma(u))$ by the shape parameter or tail index ξ and the scale parameter $\sigma(u)$. Thus, the GPD distribution groups together three distributions according to the values of the shape parameter. When $\xi = 0$, it is the **exponential distribution**, when $\xi > 0$, it is the **standard Pareto distribution**, and when $\xi < 0$, it is the **type II Pareto distribution** [39].

2.3.3 Methods for determining the extreme threshold

According to extreme value theory, a loss exceeding a given threshold is classified as an extreme loss, while a loss below this threshold is classified as an attritional loss. However, the challenge remains to determine the optimal severe loss threshold before proceeding with extreme value modeling. Indeed, a threshold must be chosen that both isolates extreme losses and retains a sufficient number of claims for modeling. A threshold that is too low allows for a sufficiently large sample size, but it negatively impacts the estimation; however, a threshold that is too high does not provide enough data for estimation. To choose the most appropriate threshold for estimation, the literature on extreme losses presents several methods for estimating the severe loss threshold, namely the Hill estimator and the mean excess function estimator.

The estimator of the mean excess function

The mean excess function describes the prediction of whether the threshold u is exceeded when an excess occurs [40]. It is defined as the ratio between the total amount of excesses relative to a certain threshold u and the total number of points exceeding this threshold u [41,42]. The estimator of the mean excess function is written as follows:

$$\hat{e}_n(u) = E[X - u \mid X > u] = \frac{\sum_{i=1}^n (X_i - u)_+}{\sum_{i=1}^n I_{X_i > u}} \tag{19}$$

where $\hat{e}_n(u)$ is the empirical estimator of the excess function. To choose the threshold u, this approach involves plotting the curve of the empirical estimator of the residual expectation of X. We choose u such that $\hat{e}_n(u)$ is approximately linear for all $x \ge u$. We seek the optimal threshold at which the curve has a linear shape [17]. If, at a certain threshold, the empirical mean excess function has a positive slope, then the data follow a Generalized Pareto distribution with a positive tail parameter [40,43]. When the empirical mean excess function has a horizontal slope, the data follow an exponential distribution. Finally, when the empirical mean excess function has a negative slope, the data follow a light-tailed distribution, i.e., a negative tail parameter [43].

Hill estimator

The Hill estimator, the most widely used in extreme value theory, was introduced in 1975 by [6,44]. It can only be used for distributions belonging to the Fréchet domain $\xi > 0$ [14]. The Hill estimator ensures a good bias-variance balance [45]. Given $X_1 \ge \ldots \ge X_n$the order statistic, n the number of observations, and k an integer less than or equal to n, the Hill estimator is given by the following formula [46,47]:

$$\hat{\xi}^{Hill}_{k,n} = \frac{1}{k}\sum_{i=1}^{k} [\ln(X_i) - \ln(X_{k+1})] \tag{20}$$

2.4 Selection of Machine learning algorithms

2.4.1 K-means

The K-Means algorithm was introduced by James MacQueen in 1967 [48]. K-Means is a classification algorithm belonging to the family of unsupervised machine learning techniques. K-Means algorithm aims to divide data into K clusters by grouping similar observations into a single cluster distinct from the other clusters [49]. To group data into the same cluster, the algorithm measures dissimilarity distances. If the distance between two observations is small, the two observations belong to the same cluster.

2.4.2 Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning tool suited to classification and regression problems [50]. The Support Vector Machine technique

© 2026 NSP
Natural Sciences Publishing Cor.

allows the construction of a classifier that aims to create a decision boundary, called a hyperplane, between two classes such that it is as far as possible from the nearest data points of each class [51].

2.4.3 K Nearest Neighbors

In 1951, E. Fix and J.L. Hodges introduced, in a technical report on nonparametric discrimination analysis and probability density estimation [52], a nonparametric classification method that became known as k-nearest neighbors [53]. K-nearest neighbors is a supervised machine learning algorithm suitable for both regression and classification problems. KNN classifies unlabeled observations by assigning them to the class of the most similar labeled examples [54,55].

2.4.4 Isolation Forest

The Isolation Forest algorithm is like the random forest algorithm but differs in that the forest's decision trees are built without any supervision, meaning without labels associated with the data. It is particularly well-suited for anomaly or outlier detection problems. The Isolation Forest algorithm randomly divides the data by randomly choosing an attribute and a division value, until no further divisions are possible.

2.4.5 Classification and Regression Tree

Classification and Regression Tree (CART) was developed by Leo Breiman in 1984 [56]. The decision tree is a supervised machine learning algorithm distinguished by its simplicity in understanding, interpreting, and explaining [57]. Furthermore, the decision tree is well-suited for solving regression and classification problems. Decision trees are hierarchical models that behave as a successive series of conditional tests [12,58]. This structure takes the form of a tree, comprising a root, internal nodes, and leaves, also called terminal nodes [59,60]. The leaves, or terminal nodes, represent the final predictions and labels resulting from the sequence of divisions performed [59].

2.4.6 Random Forest

Random Forest, a supervised machine learning method, was introduced by Leo Breiman in 2001 [61]. Random forests are an improved extension of Breiman's 1996 concept of bagging [62]. As its name suggests, a random forest is formed from multiple trees constructed simultaneously, rather than a single decision tree. What particularly distinguishes random forests is the introduction of randomness during the construction of the trees to be aggregated. This is achieved by randomly drawing a number of explanatory variables to select the optimal variable for each division. The predictions obtained from the collection of random trees are aggregated by calculating their mean or majority vote [63].

2.4.7 Boosting methods

Boosting, introduced by Freud and Schapire in 1996 [64], involves building individual models sequentially. Each new model is created by learning from the errors of the previous model. In other words, each model is an adaptive and improved version of the previous one, penalizing poorly predicted observations [65]. Among the boosting methods used in this work are XG Boost, Light GBM, Cat Boost, AdaBoost, and Gradient Boosting.

2.4.8 Bagging

Bagging, short for "Bootstrap Aggregating," one of the first ensemble classifiers [66], is an aggregation technique proposed by Leo Breiman in his article "Bagging Predictors" (1996) [62]. This method involves generating multiple versions of a predictor simultaneously and independently from a single training dataset (Bootstrap) and then aggregating them to obtain an overall predictor (Aggregating). This is achieved either by averaging the different versions for numerical values or by performing a majority vote when predicting a final class for classification [60,62,67].

2.4.9 Artificial Neural Network Algorithms

The history of artificial neural networks dates back to the 1950s [68] and the work of psychologists like Frank Rosenblatt, who introduced the perceptron algorithm [69]. The architecture of artificial neural networks was inspired by biological neural networks to mimic the human brain [67]. A neural network is composed of three distinct types of layers: an input layer, which is the first layer containing the input data; one or more hidden (intermediate) layers, which are inserted between the input and output layers; and an output layer, which is the final layer generating predictions [70]. In this work, we use three types of artificial neural network algorithms: the Autoencoder algorithm, the Deep Learning algorithm, and the Extreme Learning Machine (ELM) algorithm.

2.5 Data

To implement our proposed approach, we use an open-source database representing the fire claims portfolio [71]. This portfolio includes 27 variables and

© 2026 NSP
Natural Sciences Publishing Cor.

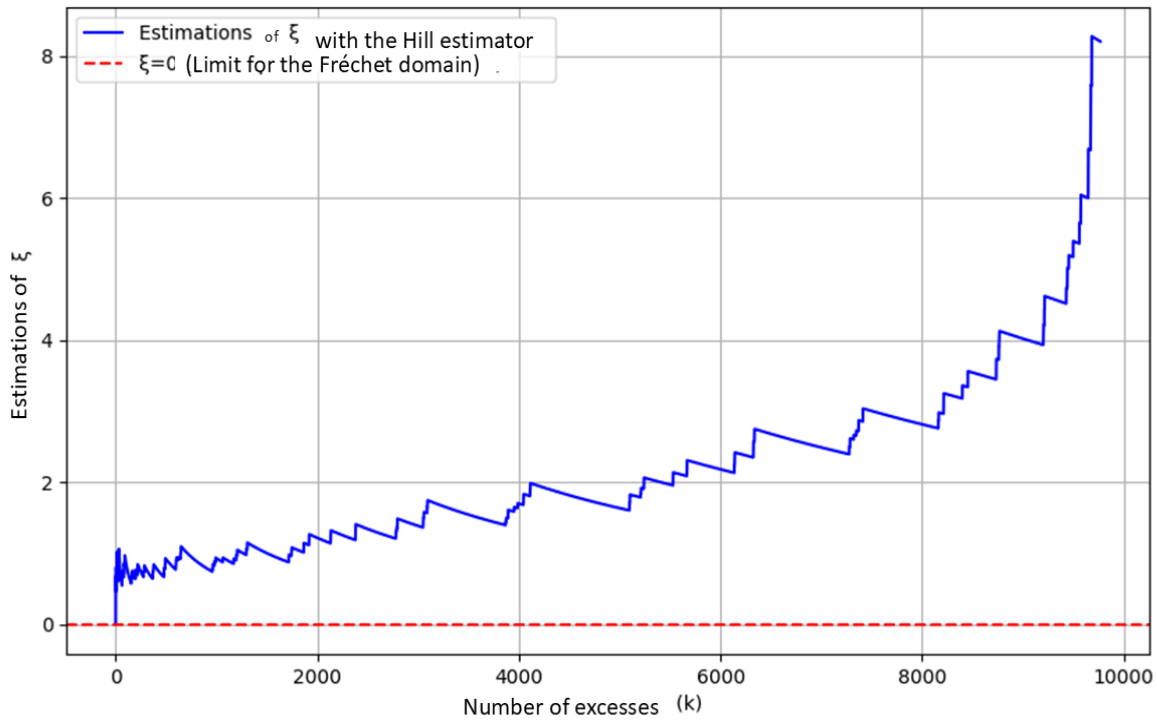

Fig. 2: Estimates of the shape parameter ξ for different excesses

11,214 observations. Among the explanatory variables to be considered are: Area of origin, Fire extent, Impact of the fire alarm system on evacuation, Operation of the fire alarm system, Presence of the fire alarm system, Ignition source, Fire control method, Possible cause, Property use, etc. The area of origin of 2 870 fires was the kitchen. 5 123 fires were confined to the object of origin and 4 084 fires were confined to a part of the room or area of origin. According to the ignition source, 1 356 fires were caused by smoking materials (e.g., cigarettes, cigars, already lit pipes). The ignition source of 2 471 fires was undetermined, and 1 916 fires were caused by stove burners. 3 719 fires occurred in multi-unit buildings with more than 12 units.

Table 2: Descriptive statistics of the variable "estimated loss in Dollars"

Estimated loss in Dollars	Value
Mean	42 943
Standard deviation	53 3936
Min – Max	0 – 50 000 000
25%	250
50%	2 500
75%	15 000
99%	500 000
Sum	481 570 578

As reported in Table 2, losses from fires range from zero to a maximum value of 50 000 000 with an average loss amount per claim of 42 943. The total amount of claims that exceed the 99th percentile of claims with a value corresponding to 500 000 weighs 49.95% of the total burden of fire claims. In other words, 1% of claims generate almost 50% of losses.

3 Results and Discussion

3.1 Results of extreme losses threshold estimation

We first applied the Hill estimator to determine the threshold for extreme claims. However, we must ensure that we are in a Fréchet domain. To do this, we estimated the tail index ξ for the different possible numbers of excesses.

Fig. 2 shows the Hill estimator plot, which plots the tail index estimates ξ as a function of the number of excesses. From this plot, we see that the tail estimator for each number of excesses is strictly positive. In this case, we can use the Hill estimator since we are in a Fréchet domain, and therefore the Hill estimator is well-defined for $\xi > 0$. However, we note that the estimator is volatile when the number of excesses is high, specifically when the number of excesses is greater than 2 000.

© 2026 NSP
Natural Sciences Publishing Cor.

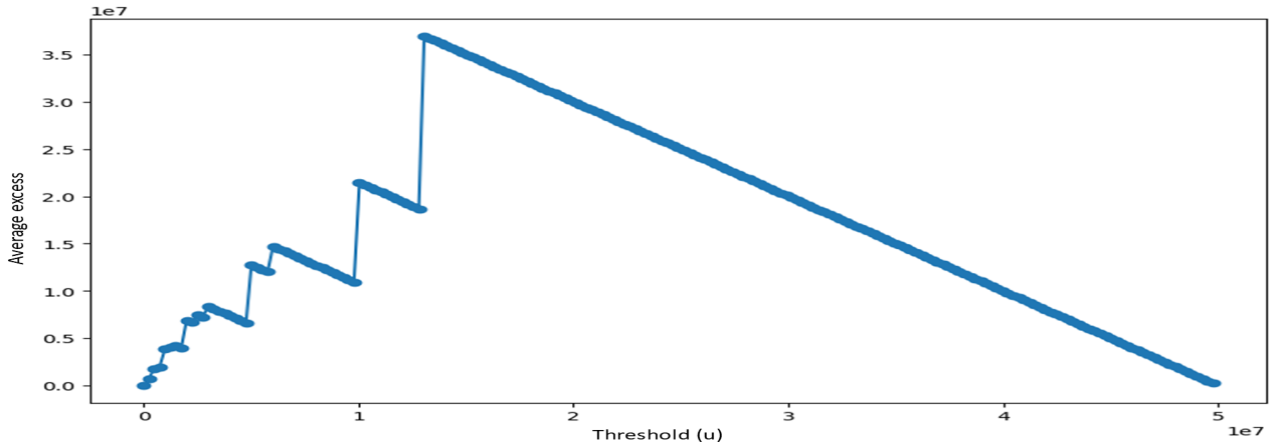

Fig. 3: The graph of the mean excess function over all data

Table 3: Hill Estimator Results

	Hill estimator
Extreme claims threshold	75 000
Number of extreme claims	1 121
Total burden of extreme losses	392 367 102

Table 4: Results of the mean excess function estimator

	The estimator of the mean excess function
Extreme claims threshold	251 256
Number of extreme claims	300
Total burden of extreme losses	284 767 000

As shown in (Table 3), the threshold for extreme losses is estimated at around 75 000 by Hill estimator. According to this estimator, a loss can be classified as extreme if the loss amount exceeds an estimated threshold of 75,000. We note that 9.49% of losses are extreme. The total burden of extreme losses represents 81.48% of the total loss burden, with an amount equal to 392 367 102. However, the burden of 90.51% of normal losses represents only 18.52% of the total loss burden.

We then determined the threshold for extreme losses using the mean excess function estimator.

According to Fig. 3, there is a change in the slope starting at the threshold of 13 000 000. Indeed, from this amount, the average excess decreases as the threshold increases. This decrease means that the number of claims exceeding this threshold is low. We note that it is not easy to determine the optimal threshold using the graph of the average excess function. To do this, we fitted the excess data to the Pareto distribution to estimate the distribution of extreme values and identify the threshold that stabilizes the average excess. For a basket of potential thresholds, we calculated the average excesses and fitted a Pareto distribution to the excess data. Finally, we chose the optimal threshold that minimizes the difference between the average excess and the Pareto distribution prediction. Fitting to a Pareto distribution allows us to choose the threshold where the model is most stable. In this sense, the threshold of extreme losses according to the estimator of the average excess function is 251 256.

Table 4 presents the results of the threshold estimation using Excess Function Estimator. According to this estimator, a loss can be classified as extreme if the amount of the loss exceeds a threshold of 251 256. We note that 2.53% of the losses are extreme, representing 59.13% of the total loss burden. However, 97.47% of normal losses represent only 40.87% of the total loss burden.

Determining the threshold with classical estimation methods from extreme value theory, including the POT Peaks Over Threshold method, has some limitations. Indeed, we note that the tail parameter estimates for the Hill estimator are unstable even if the necessary condition for the Hill estimator is satisfied (tail parameter estimates are positive). This indicates some limitations of the Hill estimator, including the need to make the assumption about the underlying distribution whose domain of attraction is Fréchet. In this case, the data may not be well modeled by a heavy-tailed distribution. Moreover, if the data do not belong to a homogeneous distribution or if there are temporal effects that cannot be captured by the Hill estimator, there will likely be an impact on the stability of the Hill estimator. Thus, the Hill estimator is not effective when the number of severe values in the data is insufficient. Therefore, the tail parameter estimates are unstable. Regarding the estimator of the mean excess function, this estimator requires an assumption about the distribution of the data. The limitation of this estimator is that the data must follow a Generalized Pareto distribution.

© 2026 NSP
Natural Sciences Publishing Cor.

Fig. 4: Silhouette score based on the number of clusters

To overcome the limitations of traditional estimators, we determined the threshold for extreme losses using the K-means algorithm as an innovative alternative to traditional methods.

First, we performed an optimization of the hyperparameters of the K-means algorithm. As illustrated in Fig. 4, the optimal number of clusters is 2 with a better silhouette score (0.9576). Indeed, the higher the silhouette score, the better the quality of the clusters formed by the K-means algorithm.

As shown in Fig. 5, the threshold for extreme losses is estimated to be around 1 421 672. According to the K-means estimator, a loss can be classified as extreme if the amount of the loss exceeds a threshold of 1 421 672.

Table 5: The results of the K-means estimator

	K-means estimator
Extreme claims threshold	1 421 672
Number of extreme claims	25
Total burden of extreme losses	70 350 000

According to Table 5, We note that 0.22% of claims are extreme, their share in the total amount being equal to 17.65% of the total claims burden. However, 99.78% of normal claims represent 82.35% of the total claims burden.

Table 6: Results of the estimation of the threshold of extreme losses

	Extreme claims threshold
Hill Estimator	75 000
Estimator of the mean excess function	251 256
K-means estimator	1 421 672

From Table 6, we conclude that the three estimators gave different results, that makes the choice a little complicated.

In Fig. 6, we observe that the Hill threshold is lower, which risks including attritional claims as extreme claims. This can distort subsequent analysis. Similarly, the mean excess function estimator yields a threshold higher than that of Hill but much lower than that of K-means, which risks considering moderate claims as extreme. However, we observe that the K-means estimator captures a reasonable number of extreme claims without risking including values that are too low or moderate. This is confirmed by the proportion of extreme claims isolated by the K-means algorithm, which does not exceed 0.5%, in line with the rare nature of extreme values.

We can conclude that the threshold estimated by the K-means algorithm is the most appropriate for the nature of our problem. Indeed, this threshold allows us to create a small set of extreme claims without including too many observations. In this sense, we use in the following the threshold of the K-means estimator which is worth 1 421 672 as the threshold for serious claims.

3.2 Results of Extreme Loss Prediction

First, we created a new claims database to which we added a new column named "grave_index." This variable has two categories: 0 if the claim is attritional or 1 if the claim is extreme and the loss amount is greater than $1 421 672.

We have a database with 28 variables and 11 214 observations, of which 25 are extreme. However, this data is highly unbalanced, which negatively impacts the prediction performance of extreme claims due to the small number of observations in the minority class. Therefore, before modeling the indicator variable, we attempted to rebalance the data using the Synthetic Minority Over-sampling Technique SMOTE [72] to address the problems associated with modeling minority observations.

To ensure the correct choice of parameters for the SMOTE algorithm, we used the Random Forest algorithm as a classifier to train the model and verify the performance of the SMOTE technique in data rebalancing. In other words, we used the SMOTE algorithm to generate samples and then applied the

© 2026 NSP
Natural Sciences Publishing Cor.

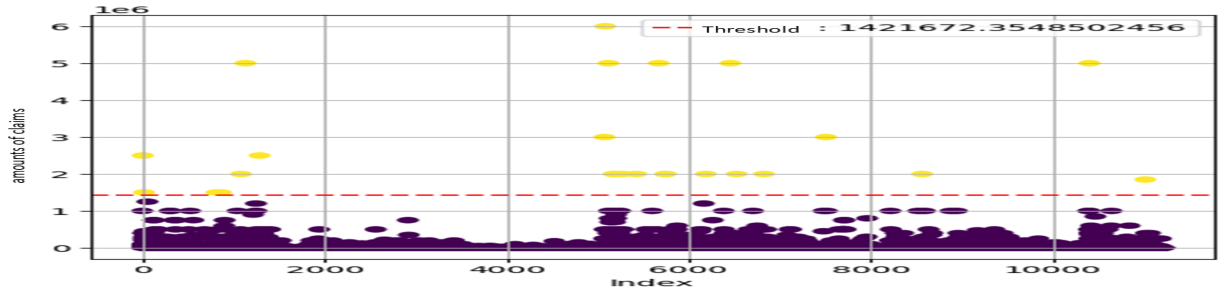

Fig. 5: Clustering result with K-means

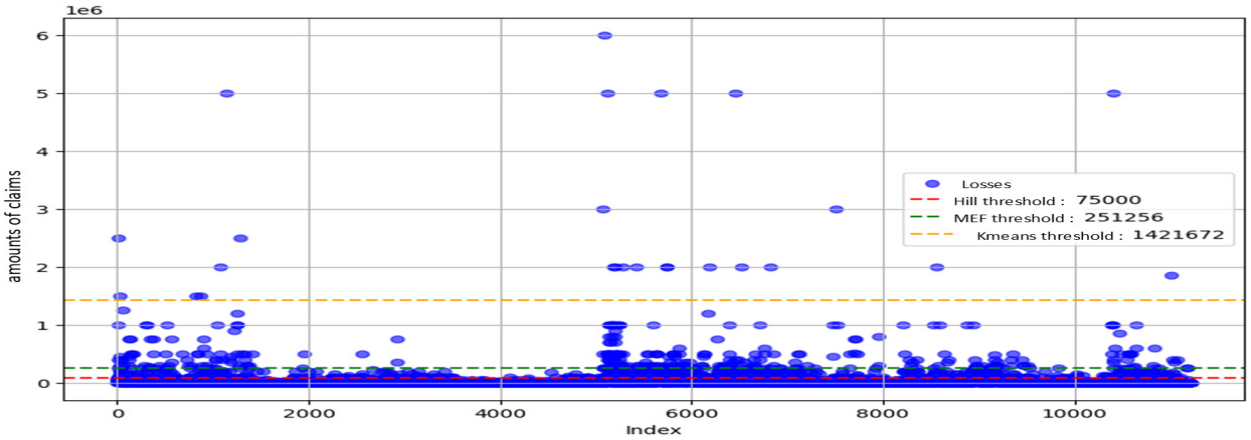

Fig. 6: Visualization of estimated thresholds (Hill, FME, K-means)

Random Forest algorithm to train it on these samples. This process is repeated during hyperparameter optimization. We optimized two parameters for SMOTE: the ”smote sampling strategy” parameter, which determines the proportion of minority classes generated relative to majority classes. The ”smote k neighbors” parameter indicates the number of k nearest neighbors used to generate new observations. We found that the optimal number of nearest neighbors is 5, and the optimal proportion of severe claims is 75% compared to normal claims in the new samples.

Fig. 7 illustrates the distribution of claims before and after using SMOTE. After using SMOTE, we have a database of 19 573 observations, including 8 388 extreme cases. We rebalanced the data by increasing the number of serious claims from 0.26% to 42.85%, compared to 57.15% of normal claims, instead of 99.74%, using the SMOTE technique.

Once the synthetic sample was generated, we modeled the occurrence of a serious claim using five machine learning algorithms: the Support Vector Machine algorithm, the K Nearest Neighbors algorithm, the Isolation Forest algorithm, the One Class SVM algorithm, and the Auto-encoder algorithm.

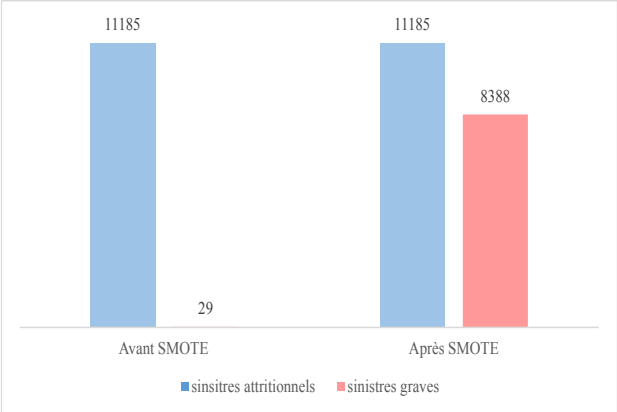

Fig. 7: Distribution of claims before and after SMOTE

We coded the categorical variables and scaled the explanatory variables using the data normalization technique. However, we end up with a database composed of 519 explanatory variables.

© 2026 NSP
Natural Sciences Publishing Cor.

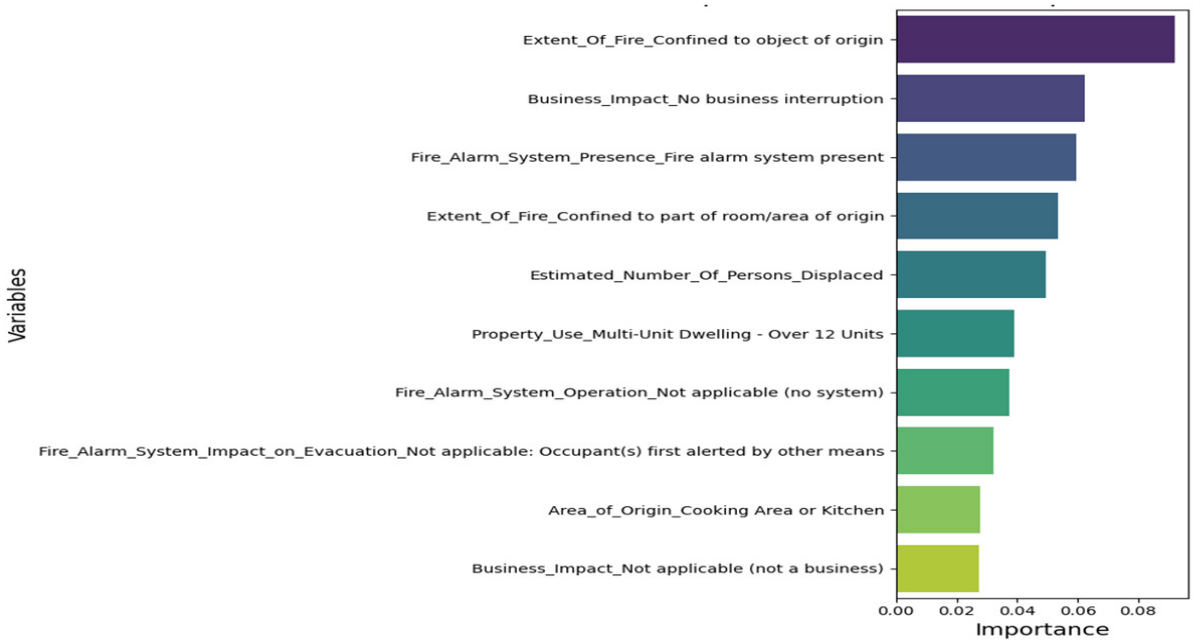

Fig. 8: The 10 most important explanatory variables

However, the use of all these explanatory variables impacts the modeling process in terms of time cost. This is why we carried out a variable selection using the Random Forest algorithm which allows us to evaluate the influence of each explanatory variable on the "grave_index" variable and to take into account the interaction between the explanatory variables. According to this algorithm, we select 50 most important explanatory variables.
Fig. 8 only shows the first 10 most important explanatory variables.

The optimization of the hyperparameters of each algorithm was carried out using the 5-fold cross-validation technique and the F1-score metric as optimization criterion. This optimization made it possible to determine the best hyperparameters to consider in the construction of each algorithm (Table 7).

We present below the results of these developed models.

From the calculated performance metrics (Table 8), we observe that the Isolation Forest model correctly predicts 47% of all observations. However, the algorithm fails to correctly detect severe losses, with a precision and recall rate of almost zero. With a specificity rate of approximately 0.82, the algorithm is able to correctly identify 82% of normal losses. However, the balance between recall and precision is not achieved, with an F1-score of almost zero. With the One-Class SVM algorithm, the model correctly predicts 59% of all observations. The algorithm successfully detects 63% of severe losses, with a precision rate of 51%. With a specificity rate of 55%, the algorithm can only correctly identify half of normal losses. Therefore, the F1-score is around 57% on average.

Regarding the SVM algorithm, the model successfully predicts 99% of all observations. Thus, the algorithm demonstrates better performance in detecting serious claims. Indeed, the model successfully detects 99% of serious claims with an accuracy rate of 99%. With a specificity rate of approximately 99%, the algorithm can correctly identify all normal claims. Therefore, the balance between Recall and Precision was satisfied with an F1-score of approximately 99%.

With the KNN algorithm, the model exhibits satisfactory performance. Indeed, the model successfully predicts 99% of all observations. In addition, the algorithm correctly detects 98% of serious claims and predicts 99% of serious claims correctly among all predicted serious claims. Similarly, the model correctly identifies 98% of normal claims. These results allowed to have a good balance between Recall and Precision with an F1-score of 99%.

Using an Autoencoder algorithm, the model correctly predicts half of all observations with an Accuracy rate of 52%. However, the algorithm fails to detect severe claims with near-zero Recall and Precision rates. With a Specificity rate of 91%, the algorithm correctly identifies all normal claims. However, the model fails to achieve a balance between Recall and Precision with a near-zero F1 score.

© 2026 NSP
Natural Sciences Publishing Cor.

Table 7: The best hyperparameters for each algorithm

Algorithm	The best hyperparameters
Isolation Forest	'contamination': 0.1, 'max_samples': 'auto', 'n_estimators': 50
One Class SVM	'kernel': 'linear', 'nu': 0.1
SVM	'C': 10, 'kernel': 'rbf'
KNN	'n_neighbors': 3, 'weights': 'distance'
Autoencoder	'input_dim': X_train_scaled.shape[1], 'encoding_neurons': [32, 16], 'decoding_neurons': [32], 'activation': ['relu', 'sigmoid'], 'optimizer': 'adam', 'loss_function': 'mse'

Table 8: Performance of extreme loss prediction models

Metric	Isolation Forest	One Class SVM	SVM	KNN	Autoencoder
Accuracy	0.4734	0.5909	0.9971	0.9918	0.5212
Recall	0.0003	0.6355	0.9988	0.9988	0.0003
Precision	0.0017	0.5189	0.9944	0.9824	0.0034
Specificity	0.8288	0.5574	0.9958	0.9865	0.9126
F1 score	0.0006	0.5713	0.9966	0.9905	0.0007

We also analyze the performance of each model based on the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) value to further our analyses.

According to Fig. 9, we observe that the ROC curve of the Isolation Forest model is very close to the upper left corner with an AUC of 0.98, indicating better model performance. We note that the ROC curve of the One-Class SVM model is close to the diagonal with a low AUC of 0.37, indicating the poor capacity of this model.

The SVM and KNN models exhibit perfect curves with an AUC of 1, demonstrating perfect performance and a better ability of these two models to correctly distinguish between severe and normal claims. However, the Autoencoder model exhibits poor performance in terms of the ROC curve and the AUC, which is only 0.07. The model is unable to distinguish severe claims from normal claims.

In general, the SVM, KNN, and Isolation Forest models are the best models, representing good performance in classifying claims into extreme and normal. However, the One-Class SVM and Autoencoder models are the least effective models in terms of distinguishing claims.

To choose the best model to use for predicting the occurrence of serious claims, we use the F1-score performance measure as a model selection criterion to ensure a balance between recall and accuracy.

As shown in Fig. 10, the Support Vector Machine algorithm exhibited the best performance among all the models used. Indeed, this model had the highest F1-score of 0.9966, followed by the KNN model with a value of 0.9905, and in third place, the One Class SVM model with an F1-score of 0.5713. However, the Isolation Forest and Autoencoder models had the lowest F1-score values, which were almost zero.

In addition, the Support Vector Machine model also demonstrated the best performance in terms of Recall, Precision, and Accuracy, with values close to 100%. The superior performance of the Support Vector Machine (SVM) model suggests that it is better suited for predicting the occurrence of extreme losses. In practice, the insurer can distinguish with high precision between attritional claims and extreme claims, thereby enabling the offer of an appropriate premium for each policyholder profile, particularly for those presenting a high risk.

3.3 Results of Extreme pure premium estimation

For the extreme loss frequency modeling framework, we used the same extreme loss modeling database with a change in the problem type. Instead of binary classification, we performed regression to determine the extreme loss frequency.

To this end, we used a basket of 12 machine learning algorithms, namely the XG Boost algorithm, the Light GBM algorithm, the Cat Boost algorithm, the Random Forest algorithm, the Deep Learning algorithm, the Support Vector Machine Regressor (SVR) algorithm, the Classification and Regression Tree (CART) algorithm, the AdaBoost algorithm, the Bagging algorithm, the Extra Trees algorithm, the Gradient Boosting algorithm, and the Extreme Learning Machine (ELM) algorithm.

We used the extreme claims database with the 50 most important variables selected using the Random Forest algorithm.

To ensure the development of optimal severe claims frequency models, we performed hyperparameter tuning for each algorithm using 3-fold cross-validation and the root mean square error (RMSE) as the optimization criterion.

Table 9 presents the best parameters obtained during

© 2026 NSP
Natural Sciences Publishing Cor.

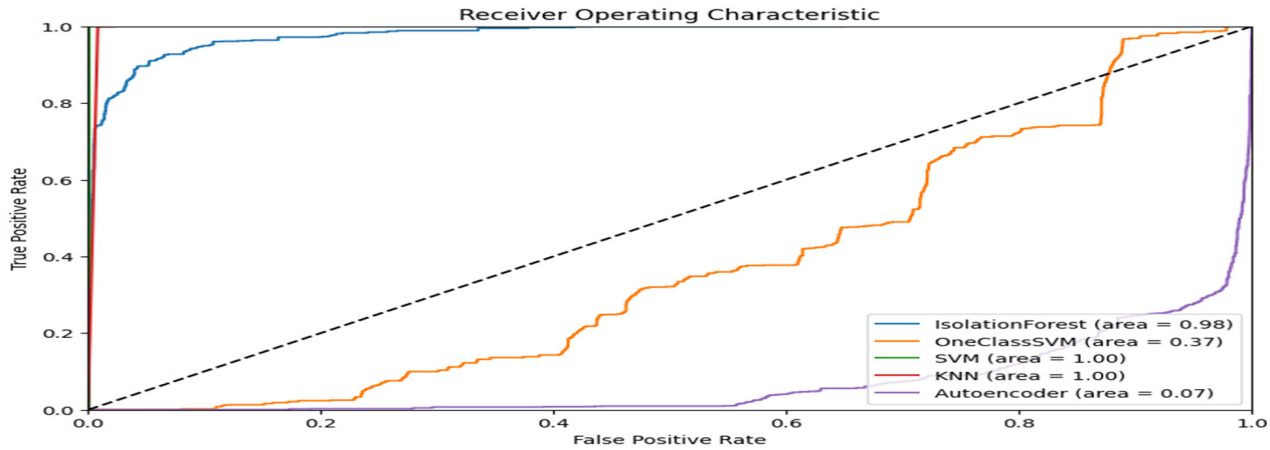

Fig. 9: ROCs of the ML algorithms used

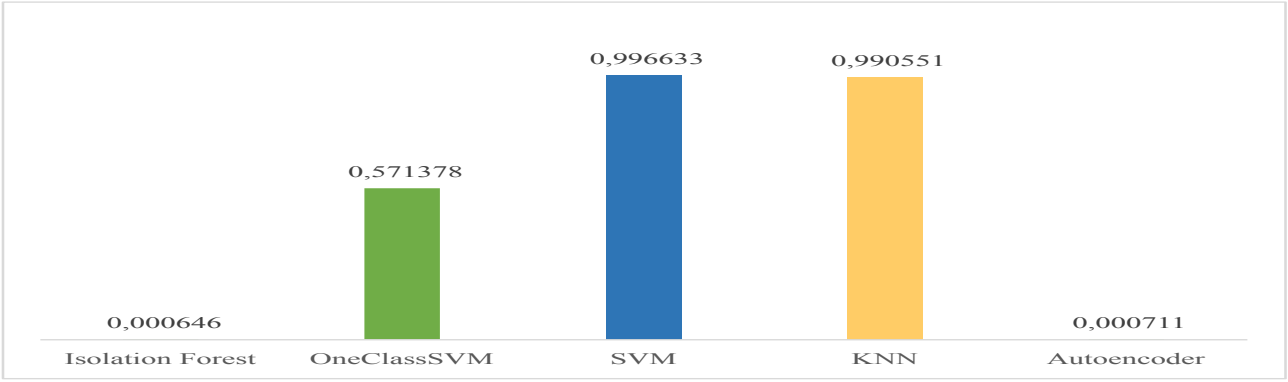

Fig. 10: F1 scores result for extreme loss prediction

hyperparameter optimization, which must be considered in the construction of each algorithm.

Below we present the performance measures of the developed models.

From Fig. 11, we observe that the AdaBoost, SVR, CART, and XG Boost models have the highest root mean square errors. These are the worst-performing models for modeling the frequency of extreme claims. However, the Light GBM, ELM, Random Forest, Deep Learning, and CAT Boost models perform best on the test set, with the lowest RMSEs around 0.04. The Cat Boost model has the lowest root mean square error among all the models in the test set, with a value of 0.043383. This model performs well in predicting the frequency of extreme claims in both the test set and the training set, where it achieved an acceptable error of 0.0166. The Cat Boost model is the best model for modeling the frequency of extreme claims.

In practical terms, the high performance of the Cat Boost model enables the insurance company to more accurately estimate the frequency of extreme claims, information required to calculate the pure premium for a policyholder with an extreme risk profile.

We move on to the presentation of the results of the extreme loss cost modeling. To model the cost of extreme losses, we created a database of extreme loss amounts. We therefore created a new variable representing the difference between the extreme loss amount and the selected extreme loss threshold. These difference amounts will be used to model the excess that the policyholder will have to pay in the event of an extreme loss. However, this new database only contains 29 observations of severe losses. Therefore, we used a technique to generate additional data, in this case the Bootstrap technique. This technique allows us to have a fairly large database. Thanks to this technique, we obtain a database composed of 10 000 observations.

A step to select the most important variables was performed using the Random Forest algorithm. This algorithm allows us to build a database with only 50 explanatory variables. We developed extreme loss cost models using the 12 machine learning algorithms used in

© 2026 NSP
Natural Sciences Publishing Cor.

Table 9: The best hyperparameters for each algorithm

Algorithm	The best hyperparameters
XG Boost	'learning_rate': 0.1, 'max_depth': 7, 'n_estimators': 500
Light GBM	'learning_rate': 0.1, 'n_estimators': 500, 'num_leaves': 50
Cat Boost	'depth': 7, 'iterations': 300, 'learning_rate': 0.1
Random Forest	'max_depth': 10, 'n_estimators': 10
Deep Learning	'hidden_layer_sizes': (100,), 'learning_rate_init': 0.01
SVR	'C': 0.1, 'kernel': 'rbf'
CART	'max_depth': 10, 'min_samples_split': 2
AdaBoost	'learning_rate': 0.1, 'n_estimators': 500
Bagging	'max_samples': 1.0, 'n_estimators': 10
Extra Trees	'max_depth': 20, 'n_estimators': 30
Gradient Boosting	'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 500
ELM	'neurons': 1000, 'activation_function': 'sigmoid'

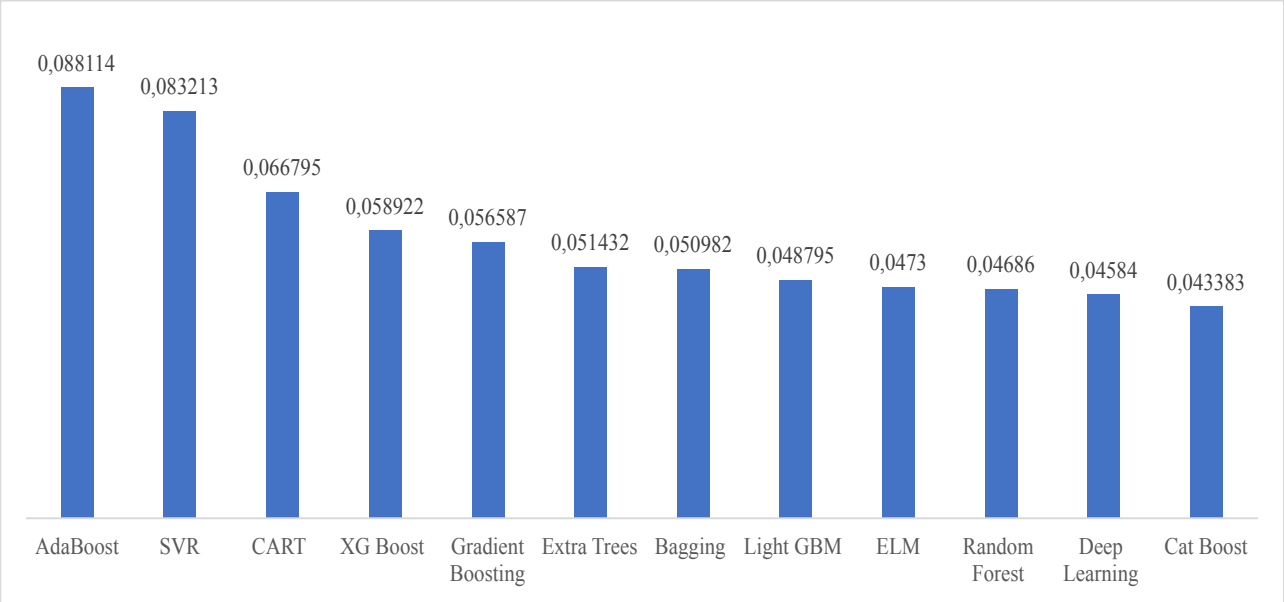

Fig. 11: RMSEs for testing extreme loss frequency prediction models

Table 10: The best hyperparameters for each algorithm

Algorithm	The best hyperparameters
XG Boost	'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 10
Light GBM	'learning_rate': 0.01, 'n_estimators': 10, 'num_leaves': 31
Cat Boost	'depth': 3, 'iterations': 100, 'learning_rate': 0.01
Random Forest	'max_depth': 10, 'n_estimators': 10
Deep Learning	'hidden_layer_sizes': (100,), 'learning_rate_init': 0.001
SVR	'C': 0.1, 'kernel': 'rbf'
CART	'max_depth': 10, 'min_samples_split': 2
AdaBoost	'learning_rate': 0.5, 'n_estimators': 10
Bagging	'max_samples': 0.5, 'n_estimators': 10
Extra Trees	'max_depth': 10, 'n_estimators': 50
Gradient Boosting	'learning_rate': 0.01, 'max_depth': 3, 'n_estimators': 10
ELM	'neurons': 1000, 'activation_function': 'sigmoid'

© 2026 NSP
Natural Sciences Publishing Cor.

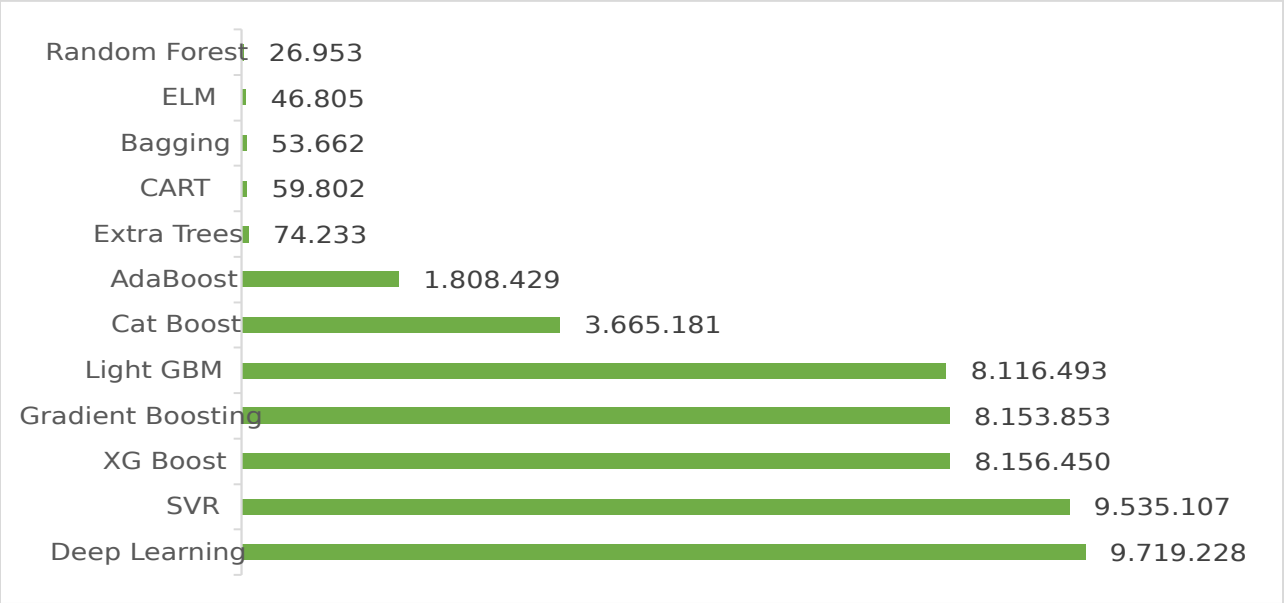

Fig. 12: RMSEs of testing extreme loss surplus prediction models

extreme loss frequency modeling. We adjusted the algorithms' hyperparameters using 3-fold cross-validation and the root mean square error (RMSE) as the optimization criterion. Table 10 presents the optimal parameters required to build the most efficient models.

Below we present the performance measures of these models in predicting extreme claim costs.

From Fig. 12, we found that Deep Learning, SVR, XG Boost, Gradient Boosting, Light GBM, CAT Boost, and AdaBoost were the worst performing models in predicting extreme loss costs, with the highest RMSE values exceeding 1 808 429. These models demonstrated limited capabilities in modeling excess extreme loss costs. However, Extra Trees, CART, Bagging, ELM, and Random Forest models demonstrated remarkable capabilities in predicting excess extreme loss costs. Indeed, these models performed best on both the test and training datasets, with the lowest RMSE values, with values below 74 233.

We observed that the Random Forest model exhibited the lowest root mean square error compared to all other machine learning algorithms used, with a value of 26 953 on the test base. This model demonstrated good performance in predicting the cost of extreme claims both on the test base and on the training base, where it achieved a minimum error of 29 161. We then used the Random Forest algorithm to model the cost of extreme claims. In fact, this model enables the insurer to predict with high precision the amount of excess claims that might be paid out if the insured presents an extreme risk profile. Using these two models, the insurer will be able to estimate the pure premium for extreme risks, the amount that the policyholder would pay if classified within the extreme-risk profile category. This premium will be calculated by multiplying the frequency estimated via CatBoost by the estimated cost of extreme claims derived from the Random Forest model. The results of this study highlight the potential of using artificial intelligence algorithms throughout the pricing process for extreme-risk insurance. Given the lack of data to model the frequency of normal claims and therefore the normal pure premium, and since modeling the pure premium for attritional claims is addressed in other works, we limited the scope of the application of the total pure premium proposed in this study to the calculation of the surplus that the insured must pay if the risks incurred are classified as extreme.

4 CONCLUSION

In this study, we proposed an innovative approach to extreme loss pricing, allowing us to determine the extreme loss threshold and estimate the pure extreme premium, using machine learning algorithms. The actuarial model developed penalizes policyholders whose risk is classified as extreme with a surplus called a pure extreme premium. This surplus was modeled by multiplying the extreme loss indicator variable by the frequency of extreme losses and the average cost of extreme losses. The results of this study demonstrate the advantages of modeling the occurrence of extreme claims using the Support Vector Machine algorithm, the cost of extreme claims using the Random Forest algorithm, and the frequency of extreme claims using the Cat Boost model. The approach adopted in this work represents a

© 2026 NSP
Natural Sciences Publishing Cor.

theoretical and practical advance over traditional methods in literature on extreme loss modeling. This work lays a cornerstone for future empirical research aimed at modeling extreme risks other than fire such as catastrophic risks that, although outside the scope of the present study, merit examination.

Acknowledgement

Not applicable.

Funding

Not applicable.

Competing interests

The authors declare that they have no competing interests

Author contributions

Authors' contributions All authors contributed equally

References

[1] U. Balasooriya and C.-K. Low, "Modeling Insurance Claims with Extreme Observations: Transformed Kernel Density and Generalized Lambda Distribution," *North American Actuarial Journal*, vol. 12, no. 2, 2008. doi:10.1080/10920277.2008.10597507.

[2] V. Brazauskas and A. Kleefeld, "Modeling Severity and Measuring Tail Risk of Norwegian Fire Claims," *North American Actuarial Journal*, 2016. doi:10.1080/10920277.2015.1062784.

[3] C. Laudagé, S. Desmettre, and J. Wenzel, "Severity modeling of extreme insurance claims for tariffication," *Insurance: Mathematics and Economics*, vol. 88, 2019. doi:10.1016/j.insmatheco.2019.06.002.

[4] Y. Wang, I. Hobæk Haff, and A. Huseby, "Modelling extreme claims via composite models and threshold selection methods," *Insurance: Mathematics and Economics*, vol. 91, pp. 257–268, 2020. doi:10.1016/j.insmatheco.2020.02.009.

[5] S. Farkas, O. Lopez, and M. Thomas, "Cyber claim analysis using Generalized Pareto regression trees with applications to insurance," *Insurance: Mathematics and Economics*, vol. 98, pp. 92–105, 2021. doi:10.1016/j.insmatheco.2021.02.009.

[6] J. Boonradsamee, U. Jaroengeratikun, and W. Bodhisuwan, "An Optimal Threshold Selection Approach for the Value at Risk of the Extreme Events," *Lobachevskii Journal of Mathematics*, vol. 43, no. 9, pp. 2397–2410, 2022. doi:10.1134/S1995080222120071.

[7] W. Xu, Y. Hou, and D. Li, "Prediction of Extremal Expectile Based on Regression Models With Heteroscedastic Extremes," *Journal of Business & Economic Statistics*, vol. 40, no. 2, pp. 522–536, 2022.

[8] I. C. Sabban, O. Lopez, and Y. Mercuzot, "Automatic analysis of insurance reports through deep neural networks to identify severe claims," *Annals of Actuarial Science*, vol. 16, no. 1, pp. 42–67, 2022. doi:10.1017/S174849952100004X.

[9] J. V. F. Carvalho and L. H. A. Oliveira, "We are Living on the Edge: Managing Extreme-Severity Claims Using Extreme Value Theory," *Brazilian Business Review*, vol. 21, no. 3, 2024. doi:10.15728/bbr.2022.1245.en.

[10] D. Ding, M. Zhang, X. Pan, M. Yang, and X. He, "Modeling Extreme Events in Time Series Prediction," in *Proceedings of ACM*, pp. 1114–1122, 2019. doi:10.1145/3292500.3330896.

[11] Y. Kouach, A. E. Attar, and M. E. Hachloufi, "Auto insurance fraud detection using unsupervised learning," *Alternatives Managément Economiques*, vol. 4, no. 4, 2022. doi:10.48374/IMIST.PRSM/ame-v4i4.35537.

[12] Y. Kouach, A. El Attar, and M. El Hachloufi, "Automobile Takaful Pricing by Machine Learning Algorithms," *Recherche Appliquée en Finance Islamique*, vol. 6, no. 2, p. 18, 2022. doi:10.48394/IMIST.PRSM/rafi-v6i2.32630.

[13] K. Yassine, E. A. Abderrahim, and E. H. Mostafa, "Retakaful Contributions Model Using Machine Learning Techniques," *Journal of Islamic Monetary Economics and Finance*, vol. 9, no. 3, pp. 511–532, 2023.

[14] O. Bahi, "Théorie des Valeurs Extrêmes : Application au Calcul de Risques," Ph.D. thesis, Université des Frères Mentouri Constantine 1, 2019. doi:10.13140/RG.2.2.27147.64807.

[15] J. V. F. Carvalho and L. H. A. Oliveira, "We are Living on the Edge: Managing Extreme-Severity Claims Using Extreme Value Theory," *Brazilian Business Review*, vol. 21, 2024. doi:10.15728/bbr.2022.1245.en.

[16] H. M. Barakat, O. M. Khaled, and E.-S. M. Nigm, *Statistical Techniques for Modelling Extreme Value Data and Related Applications*. Cambridge Scholars Publishing, 2019.

[17] C. Tremblay, "Prédire les sinistres graves en assurance : les apports de l'apprentissage statistique aux modèles linéaires," Master thesis, Université de Paris, ENSAE, 2017.

[18] C. L. Smith, "Representing external hazard initiating events using a Bayesian approach and a generalized extreme value model," *Reliability Engineering & System Safety*, vol. 193, 2020. doi:10.1016/j.ress.2019.106650.

[19] A. Essadik and B. Achchab, "L'analyse des sinistres extremes en Takaful," *Revue de Gestion et d'Économie*, vol. 5, no. 1–2, 2017.

[20] B. Dai, Y. Xia, and Q. Li, "An extreme value prediction method based on clustering algorithm," *Reliability Engineering & System Safety*, vol. 222, p.108442, 2022. doi:10.1016/j.ress.2022.108442.

[21] R. A. Fisher and L. H. C. Tippett, "Limiting forms of the frequency distribution of the largest or smallest member of a sample," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol.24, no.2, 1928. doi:10.1017/S0305004100015681.

[22] E. J. Gumbel, "Les valeurs extrêmes des distributions statistiques," *Annales de l'Institut Henri Poincaré*, vol.5, no.2, 1935.

[23] B. Gnedenko, "Sur la distribution limite du terme maximum d'une série aléatoire," *Annals of Mathematics*, vol.44, no.3, 1943. doi:10.2307/1968974.

[24] M. Kereszturi, *Assessing and Modelling Extremal Dependence in Spatial Extremes*. Lancaster University, 2016.

© 2026 NSP
Natural Sciences Publishing Cor.

[25] T. Roncalli, *Handbook of Financial Risk Management*. New York: Chapman and Hall/CRC, 2020. doi:10.1201/9781315144597.
[26] K. Müller and W.-D. Richter, "Modelling extremal data," *Journal of Statistical Computation and Simulation*, vol. 87, no. 5, pp. 933–955, 2017. doi:10.1080/00949655.2016.1238089.
[27] C. Murphy-Barltrop, "Novel methodology for the estimation of extremal bivariate return curves," Ph.D. thesis, Lancaster University, 2023. doi:10.17635/lancaster/thesis/1996.
[28] M. Fréchet, "Sur la loi de Probabilité de l'écart Maximum," *Annales de la Société Polonaise de Mathématique*, vol. 6, 1927.
[29] W. Weibull, "A Statistical Distribution Function of Wide Applicability," *Journal of Applied Mechanics*, 1951.
[30] A. F. Jenkinson, "The frequency distribution of the annual maximum (or minimum) values of meteorological elements," *Quarterly Journal of the Royal Meteorological Society*, vol. 81, no. 348, 1955.
[31] F. Longin, "La théorie des valeurs extrêmes : présentation et premières applications en finance," *Journal de la Société de Statistique de Paris*, vol. 136, no. 1, 1995.
[32] K. Cooray, "Generalized Gumbel distribution," *Journal of Applied Statistics*, vol. 37, no. 1, pp. 171–179, 2010. doi:10.1080/02664760802698995.
[33] J. Beirlant, T. de Wet, and Y. Goegebeur, "A goodness-of-fit statistic for Pareto-type behaviour," *Journal of Computational and Applied Mathematics*, vol. 186, no. 1, pp. 99–116, 2006. doi:10.1016/j.cam.2005.01.036.
[34] J. I. Pickands, "Statistical Inference Using Extreme Order Statistics," *The Annals of Statistics*, vol. 3, no. 1, 1975. doi:10.1214/aos/1176343003.
[35] S. Savarre and B. Payre, "Charge ultime nette de réassurance en RC corporelle : 2 modèles stochastiques pour les flottes automobiles," *Centre des Études Actuarielles*, 2012.
[36] C. Butaci, S. Dzitac, I. Dzitac, and G. Bologa, "Prudent decisions to estimate the risk of loss in insurance," *Technological and Economic Development of Economy*, vol. 23, no. 2, pp. 428–440, 2017. doi:10.3846/20294913.2017.1285365.
[37] A. A. Balkema and L. de Haan, "Residual Life Time at Great Age," *The Annals of Probability*, vol. 2, no. 5, 1974. doi:10.1214/aop/1176996548.
[38] M. Tiene and D. Barro, "Modelling Climate Change in a Stochastic Extreme Values Framework," *Contemporary Mathematics*. Available online: https://ojs.wiserpub.com/index.php/CM/article/view/3726.
[39] Y. Li et al., "Weather Regimes of Extreme Wind Speed Events in Xinjiang: A 10–30 Year Return Period Analysis," *Atmosphere*, vol. 16, no. 10, p. 1117, 2025. doi:10.3390/atmos16101117.
[40] W. D. Walls and Wei Zhang, "Using Extreme Value Theory to Model Electricity Price Risk with an Application to the Alberta Power Market," *Energy Exploration & Exploitation*, vol. 23, no. 5, pp. 375–403, 2005. doi:10.1260/014459805775992690.
[41] S. Derkaoui, "Modélisation de la sinistralité grave dans le cadre des revalorisations de primes du partenariat," Master thesis, Institut des Actuaires, 2021.
[42] S. Ghaddab, M. Kacem, C. de Peretti, and L. Belkacem, "Extreme severity modeling using a GLM-GPD combination: application to an excess of loss reinsurance treaty," *Empirical Economics*, vol. 65, no. 3, pp. 1105–1127, 2023. doi:10.1007/s00181-023-02371-4.
[43] T. Mikosch (Ed.), "The Total Claim Amount," in *Non-Life Insurance Mathematics: An Introduction with Stochastic Processes*. Berlin, Heidelberg: Springer, 2004, pp. 77–154. doi:10.1007/3-540-44889-6_3.
[44] B. M. Hill, "A Simple General Approach to Inference About the Tail of a Distribution," *The Annals of Statistics*, vol. 3, no. 5, 1975. doi:10.1214/aos/1176343247.
[45] H. Drees, S. Resnick, and L. de Haan, "How to make a Hill plot," *The Annals of Statistics*, vol. 28, no. 1, 2000.
[46] F. Longin (Ed.), *Extreme Events in Finance: A Handbook of Extreme Value Theory and its Applications*. Hoboken, New Jersey: Wiley, 2017.
[47] N. M. Markovich, A. Undheim, and P. J. Emstad, "Classification of slice-based VBR video traffic and estimation of link loss by exceedance," *Computer Networks*, vol. 53, no. 7, pp. 1137–1153, 2009. doi:10.1016/j.comnet.2008.12.020.
[48] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 5.1, University of California Press, 1967.
[49] J. A. Hartigan and M. A. Wong, "A k-means clustering algorithm," *Applied Statistics*, vol. 28, no. 1, 1979.
[50] F. Ndikumana et al., "Machine learning-based predictive modelling of mental health in Rwandan Youth," *Scientific Reports*, vol. 15, no. 1, p. 16032, 2025. doi:10.1038/s41598-025-00519-z.
[51] S. Huang, N. Cai, P. P. Pacheco, S. Narrandes, Y. Wang, and W. Xu, "Applications of Support Vector Machine (SVM) Learning in Cancer Genomics," *Cancer Genomics & Proteomics*, vol. 15, no. 1, 2018.
[52] E. Fix and J. L. Hodges, "Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties," *International Statistical Review*, vol. 57, no. 3, 1989. doi:10.2307/1403797.
[53] L. E. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, 2009.
[54] Z. Zhang, "Introduction to machine learning: k-nearest neighbors," *Annals of Translational Medicine*, vol. 4, no. 11, 2016.
[55] P. Dini, A. Elhanashi, A. Begni, S. Saponara, Q. Zheng, and K. Gasmi, "Overview on Intrusion Detection Systems Design Exploiting Machine Learning for Networking Cybersecurity," *Applied Sciences*, vol. 13, no. 13, p. 7507, 2023. doi:10.3390/app13137507.
[56] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Wadsworth Statistics/Probability Series. Belmont, CA: Wadsworth Advanced Books and Software, 1984.
[57] J. Unpingco, *Python for Probability, Statistics, and Machine Learning*, 2nd ed. Springer Cham, 2019.
[58] A. Chloé-Agathe, *Introduction au Machine Learning*. France: Dunod, 2019.
[59] I. Hull, *Machine Learning for Economics and Finance in TensorFlow 2: Deep Learning Models for Research and Industry*. 1st ed. Apress, 2021.
[60] V. Vaibhav, *Supervised Learning with Python: Concepts and Practical Implementation Using Python*. Apress, 2020.

© 2026 NSP
Natural Sciences Publishing Cor.

[61] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, 2001. doi:10.1023/A:1010933404324.
[62] L. Breiman, "Bagging Predictors," *Machine Learning*, vol. 24, no. 2, 1996. doi:10.1007/BF00058655.
[63] R. P. Fotia Santsa, "Méthodes d'apprentissage statistique appliquées à la tarification non-vie : Tarification Automobile," Master thesis, Université des Montagnes, Faculté des Sciences et Techniques, Cameroun, 2018.
[64] Y. Freund and R. E. Schapire, "Experiments with a New Boosting Algorithm," in *Machine Learning: Proceedings of the Thirteenth International Conference*, United States, North America, 1996.
[65] E. S. Nana Njoya, "Prédiction des comportements de rachat en épargne individuelle : une approche machine learning," Master thesis, Université de Paris, ENSAE, 2016.
[66] J. Kozak, *Decision Tree and Ensemble Learning Based on Ant Colony Optimization*. Studies in Computational Intelligence. Springer Cham, 2019. doi:10.1007/978-3-319-93752-6.
[67] K. Ramasubramanian and A. Singh, *Machine Learning Using R: With Time Series and Industry-Based Use Cases in R*, 2nd ed. Apress, 2019.
[68] F. V. M. Nunes, L. M. P. Behrens, R. D. Weimer, G. F. Gonçalves, G. da Silva Fernandes, and M. Dorn, "Deep learning methods and applications in single-cell multimodal data integration," *Molecular Omics*, vol. 21, no. 6, pp. 545–565, 2025. doi:10.1039/D5MO00062A.
[69] F. Rosenblatt, "The Perceptron - a perceiving and recognizing automation," Cornell Aeronautical Laboratory, 1957.
[70] T. Roncalli, *Handbook of Financial Risk Management*. New York: Chapman and Hall/CRC, 2020. doi:10.1201/9781315144597.
[71] R. Namdari, "Fire Incidents," Kaggle Dataset, 2016.
[72] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, 2002. doi:10.1613/jair.953.

Yassine Kouach holds a PhD in Economics and Management Sciences from Hassan II University of Casablanca, Morocco. Since 2025, he has been serving as an Research Professor of Finance at ENCG El Jadida, Chouaib Doukkali University of El Jadida, Morocco. His research interests include: Actuarial Science, Insurance, Finance, Risk Management, and Machine Learning.

Abderrahim EL Attar holds PhD in Statistics and Actuarial Mathematics, at Mohammed V University, Rabat, Morocco. Since 2018 up to present time he has been working as an Research Professor in Statistics and Actuarial Mathematics at the FSJES, Ain Sebaâ, Hassan II University, Morocco. Research area: Statistics, Machine learning, Actuarial science and risk management.

© 2026 NSP
Natural Sciences Publishing Cor.