Earthline Journal of Mathematical Sciences
E-ISSN: 2581-8147
Volume 15, Number 6, 2025, Pages 1021-1036
https://doi.org/10.34198/ejms.15625.10211036

Threshold Selection for Peak Over Threshold Models Using Logistic Regression

Temitope Comfort Iroko[1,2,*], Iliyasu Tukur[3,4] and Victor Adeyanju[5]

[1] Department of Mathematics, University of Wisconsin-Milwaukee, USA
e-mail: tciroko@uwm.edu

[2] African Institute for Mathematical Sciences, Limbe, Cameroon

[3] Helpman Development Institute, Abuja, Nigeria
e-mail: iliyasutukur01@gmail.com

[4] Department of Applied Mathematics, Wroclaw University of Science and Technology, Poland

[5] Department of Mathematics, Tai Solarin University of Education, Nigeria
e-mail: victoradeyanju036@gmail.com

Abstract

Threshold selection remains a critical challenge in the application of Extreme Value Theory (EVT), particularly in actuarial science, where accurate modeling of extreme insurance claims is vital for solvency, capital adequacy, and reinsurance pricing. Traditional graphical tools, such as the mean residual life (MRL) plot, are highly dependent on visual interpretation and expert judgment, which limits reproducibility and consistency between practitioners. This paper proposes a machine learning approach using logistic regression to assign extremeness probabilities to insurance claims. The model is trained on labels generated from an initial quantile-based rule, classifying claims above the 90th percentile as extreme. The optimal threshold is determined as the smallest amount of claim with predicted probability exceeding a predefined cut-off (e.g., 90%). This probability-based rule provides an objective alternative to visual diagnostics and aligns well with classical EVT tools such as the MRL plot. After identifying exceedances above the selected threshold, a Generalized Pareto Distribution (GPD) is fitted using maximum likelihood estimation. Tail-based risk measures, including Value at Risk (VaR) and Expected Shortfall (ES), are then computed at the 99% confidence interval to quantify the severity of potential extreme losses. The proposed framework is interpretable, reproducible, and readily applicable in actuarial workflows, offering a more consistent and automated solution for tail risk modeling.

Received: July 9, 2025; Accepted: August 16, 2025; Published: September 11, 2025
2020 Mathematics Subject Classification: 62G32, 62P05, 62J12.
Keywords and phrases: extreme value theory, threshold selection, logistic regression, generalized Pareto distribution, tail modeling.
*Corresponding author
Copyright © 2025 the Authors

1 Introduction

Extreme Value Theory (EVT) is a statistical framework for modeling rare but severe events in the tails of probability distributions. In actuarial practice, EVT is essential for modeling large insurance claims, setting retention limits, and evaluating solvency risk [5, 7]. It is particularly valuable in excess-of-loss reinsurance, where accurate modeling of extreme losses directly influences pricing, capital requirements, and regulatory compliance under frameworks such as Solvency II.

A widely used EVT technique is the Peaks Over Threshold (POT) method, which models exceedances above a chosen threshold using the Generalized Pareto Distribution (GPD) [6, 11]. However, the effectiveness of the POT method depends critically on the selection of an appropriate threshold. When a threshold is too low, it violates the assumptions of the GPD and introduces bias, while a high threshold increases variance due to fewer exceedances [12]. This bias-variance trade-off has a significant impact on the stability of tail-based estimates, such as Value at Risk (VaR) and Expected Shortfall (ES).

Traditional threshold selection relies heavily on graphical diagnostics, such as the Mean Residual Life (MRL) plot or parameter stability plots [5, 10]. Although these tools are informative, they require expert judgment to visually identify threshold stability, which introduces subjectivity and limits reproducibility. This subjectivity poses challenges in actuarial practice, where consistent and reproducible modeling decisions are essential. Consequently, there is growing interest in objective, automated, and data-driven approaches to threshold selection.

This paper proposes a threshold selection framework using a supervised machine learning approach based on logistic regression within the POT method. The model estimates the probability that each claim is extreme, and the threshold is chosen as the point where this probability exceeds a predefined probability cut-off point. This probability-based approach reduces subjectivity while maintaining consistency with traditional graphical diagnostics. We apply the method to medical insurance claim data and demonstrate its alignment with classical EVT tools. The proposed framework enhances reproducibility and is well-suited for integration into actuarial workflows, particularly in regulatory reporting, capital modeling, and reinsurance pricing, where consistent tail risk estimation is crucial.

2 Literature Review

EVT is widely applied in fields such as finance, insurance, hydrology, and engineering to quantify the probability of rare and high-impact events. In actuarial science, EVT plays a crucial role in modeling the severity of extreme insurance claims and assessing tail risk for solvency and reinsurance pricing purposes. A central challenge in the practical application of EVT is selecting an appropriate threshold when fitting

the GPD, particularly under the POT framework.

2.1 Theoretical Foundations of EVT

EVT is underpinned by the Fisher-Tippett-Gnedenko theorem, which states that the distribution of block maxima converges to one of three limiting distributions: Gumbel, Fréchet, or Weibull, regardless of the parent distribution [5]. This convergence underlies the Block Maxima (BM) method. However, the POT method has become the preferred technique in many applications due to its more efficient use of data [11]. The POT method models exceedances over an optimal threshold using the GPD, which provides flexibility for modeling heavy tails and is well-suited for insurance data. However, the effectiveness of the POT method is highly sensitive to the choice of threshold, making threshold selection an important technique.

2.2 Challenges in Threshold Selection

Although GPD offers a robust framework for modeling extremes, the performance of POT methods depends on the choice of threshold. Selecting a threshold that is too low can violate the asymptotic assumptions of the model, while selecting one that is too high reduces the number of exceedances, increases variance, and reduces the stability of parameter estimates. This bias-variance tradeoff has significant implications for the accuracy of tail risk measures.

In practice, threshold selection often relies on visual diagnostics such as the MRL plot or parameter stability plots [5, 10, 12]. Although widely used, these tools require subjective interpretation and expert judgment to identify regions of stability. This subjectivity limits reproducibility and can lead to inconsistencies across practitioners. Therefore, there is a need for automated approaches to threshold selection.

2.3 Recent Advances on Threshold Selection

To overcome the limitation of graphical diagnostics, recent research has introduced a range of threshold selection methods. One class of approaches emphasizes model-based tools, including penalized likelihood methods and ordered goodness-of-fit tests with false discovery rate adjustments [2]. Automated approaches to threshold selection, including clustering and mixture models, have also been proposed for greater objectivity [15]. While penalized likelihood methods can stabilize estimation by balancing model fit and complexity, they require careful tuning of penalty terms and may be computationally intensive. Clustering and mixture models provide a data-driven partition of exceedances and improve objectivity, but often at the cost of higher model complexity and reduced interpretability. In contrast, the logistic regression

framework proposed in this study offers a more straightforward and more interpretable alternative. It provides reproducible thresholds without extensive tuning, making it particularly suitable for actuarial workflows where clarity and interpretability are critical.

Another research stream has explored machine learning models for classifying and predicting rare events. Logistic regression, for example, is widely used for binary classification under class imbalance and is often used to estimate the probability of extremeness [8]. However, despite its wide use in classification, logistic regression has not been explicitly explored for threshold selection within the POT framework. Recent studies have also emphasized the need for reproducible threshold selection frameworks in financial modeling [9].

More advanced approaches including ensemble methods such as extremal random forests, which capture complex interactions in the tail [16], and high-dimensional quantile regression methods designed to handle covariates in extreme value modeling [14]. Although promising, these techniques often require complex tuning, which limits their accessibility in actuarial workflows. As such, there is a practical need for simpler, interpretable solutions that maintain statistical rigour without compromising usability.

2.4 Contribution of Study

Despite recent advances, most threshold selection techniques require complex model tuning or depend on visual diagnostics, limiting their reproducibility and accessibility in practice. This study introduces a simple yet effective threshold selection approach based on logistic regression, offering a data-driven and interpretable alternative to identify extreme insurance claims.

The proposed method assigns the extremeness probabilities to observations and defines the threshold as the smallest claim value whose predicted probability exceeds a specified cut-off. This probability based rule reduces subjectivity and aligns with classical EVT tools, such as MRL plot, while providing a reproducible solution.

3 Extreme Value Analysis

Extreme value analysis models extreme events, which are rare large deviations from usual events with significant impacts. It is based on the asymptotic behavior of the observed extremes. This is important in modeling insurance claims where unusual events occur, leading to severe loss of lives and properties, exposing an insurer to risk. The definition of extreme value theory was adapted from [4].

3.0.1 Classical EVT via Block Maxima

Let $X_1, ..., X_n$ be n independent and identically distributed random variables with distribution function $F(x) = P\{X \leq x\}$. The maximum order statistic is defined as:

$$M_n = max\{X_1, ..., X_n\}. \tag{1}$$

As described by [5], the distribution of M_n is one of the three extreme value distributions: Gumbel Fréchet, or Weibull.

3.1 Method of Block Maxima

One of the approaches to studying extremes is the block maxima method, which considers the distribution of the maximum-order statistics defined in (1). In this method, the sample of extreme values is obtained by selecting the maximum (in some cases, the minimum) values observed in each block. This technique entails the division of independent and identically distributed random variables into non-overlapping blocks of equal length and fitting the GEV distribution to the set of maxima (minima) resulting from the blocks. The choice of block size is the only sensitive stage in the block maxima approach. Hence, care must be taken when choosing an appropriate block size because choosing a small block size will result in bias estimation. Blocks are typically one year long (365 daily observations per block) or occasionally a season long. The generalized extreme value (GEV) distribution is appropriate for block maxima when the blocks are sufficiently large.

The cumulative distribution of the GEV distribution is given as follows.

$$F(x;\mu,\sigma,\xi) = \begin{cases} \exp\left[-\left(1+\xi(\frac{x-\mu}{\sigma})\right)^{-\frac{1}{\xi}}\right], & \xi \neq 0 \\ \\ \exp\left[-\exp\left(-(\frac{x-\mu}{\sigma})\right)\right], & \xi = 0. \end{cases} \tag{2}$$

It is defined on the set$\left\{x : 1+\xi\left(\frac{x-\mu}{\sigma}\right) > 0\right\}$, where $\sigma > 0, \mu \in \mathbb{R}$, and $\xi \in \mathbb{R}$. The GEV distribution is parametrized with a shape parameter ξ, a scale parameter σ, and a location parameter μ. When $\xi = 0, > 0, < 0$, then it is referred to as the Gumbel, Fréchet, and Weibull distribution respectively.

While the block maxima method has been widely applied, this study adopts the POT framework, which models exceedances over a chosen threshold using the GPD. This approach is particularly well-suited for modeling insurance and financial data, as it leverages more information by accounting for all exceedances above the threshold rather than relying solely on a single extreme value per block (e.g., annual maxima). Additionally, the POT method avoids years where the block maximum may not adequately represent the

upper tail behavior of the underlying distribution, thereby improving the accuracy and robustness of tail modeling.

3.2 Peak Over Threshold (POT)

The POT method is used to represent the behavior of exceedance above a chosen threshold. It requires careful selection of the threshold as described in [17] and discussed in [2].

Let X be a claim amount with distribution function F, $x_F = \sup\{x; F(x) < 1\}$ be the maximum claim. For all $u < x_F$, the distribution function of the exceedance above the threshold u is given as:

$$F_u(x) = P\{X - u \leq x | X > u\}, \quad x \geq 0.$$

This method involves choosing a suitable threshold u, which is neither too high nor too low. The claim amount above the threshold is assumed to follow the Generalized Pareto Distribution, allowing the modeling of tail risks effectively. The GPD has been extensively used in insurance for modeling heavy-tailed claim distributions due to its flexibility and theoretical justification for threshold exceedances [1]. POT retains more data improving model efficiency as compared to the block maxima method. According to [5], the cumulative distribution of the GP distribution is defined as:

$$F(x) = \begin{cases} 1 - \left(1 + \xi\left(\dfrac{x-u}{\sigma}\right)\right)^{-1/\xi}, & \text{if } \xi \neq 0 \\ 1 - \exp\left(-\left(\dfrac{x-u}{\sigma}\right)\right), & \text{if } \xi = 0 \end{cases} \tag{3}$$

where $x \geq 0$ if $\xi \geq 0$, and $0 \leq x \leq -\sigma/\xi$ if $\xi < 0$, σ and ξ are the scale and shape parameter. The GPD includes three types of distributions:

- $\xi = 0$, Exponential distribution
- $\xi = 1$, Standard Fréchet distribution
- $\xi = -1$, Uniform distribution

3.3 The Mean Residual Life Plot

The Mean Residual Life Plot (MRLP) is a graphical tool widely used for threshold selection in the POT framework as proposed by [6]. Let X be a random variable, the excess above a threshold u, defined as

$X - u|X > u$, also follows the GPD with the same shape parameter ξ, but a different scale parameter that increases linearly with u:

$$\sigma_u = \sigma + \xi(u - u^*).$$

The mean excess is given as:

$$e_i = \mathbb{E}\left[X - u \mid X > u\right] = \frac{\sigma_u}{1 - \xi} \quad \text{for } \xi < 1. \tag{4}$$

3.4 Automated Threshold Selection Using Logistic Regression

To complement graphical threshold diagnostics, we apply logistic regression as a supervised machine learning classifier to distinguish between extreme and non-extreme claims. As part of our methodology, we generated the binary labels by classifying claims as extreme if they exceeded the $90th$ percentile of the empirical distribution. We define the target variable Y_i as:

$$Y_i = \begin{cases} 1 & \text{if } X_i > u_{\text{ini}} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where X_i is the claim amount and u_{ini} is the initial threshold chosen using empirical quantile (e.g the 90th percentile).

The logistic regression model is :

$$\pi(X_i) = \mathbb{E}(Y \mid X_i) = \frac{e^{\beta_0 + \beta_1 X_i}}{1 + e^{\beta_0 + \beta_1 X_i}} \tag{6}$$

and the logit form:

$$\ln\left(\frac{\pi(X_i)}{1 - \pi(X_i)}\right) = \beta_0 + \beta_1 X_i. \tag{7}$$

Model parameters β_0, β_1 are estimated using maximum likelihood estimation. The resulting predicted probability $\hat{p}_i$ reflects the likelihood that a claim is extreme.

Logistic regression is selected in this study due to its interpretability, simplicity, and performance in binary classification with imbalanced data. Compared to more complex models such as neural networks or ensemble methods, logistic regression provides easily explainable coefficients. It requires minimal tuning, making it suitable for actuarial applications where transparency and reproducibility are critical.

3.4.1 Threshold Selection via Model Probability

We defined the threshold u_{ML} as the smallest claim value whose predicted probability exceeds a specified threshold τ, typically 0.9:

$$u_{\text{ML}} = \min\{X_i : \hat{p}_i \geq \tau\}. \tag{8}$$

This threshold is then used to define exceedances for fitting the Generalized Pareto Distribution.

3.5 Parameter Estimation

The selected threshold u_{ML} is simply denoted as u from henceforth for notational convenience. The parameters of the GPD are estimated using the maximum likelihood method. Let $x_1, x_2, ..., x_n$ denote the exceedances over the selected threshold u, where n_u is the number of exceedances. The log likelihood function for the GPD, as given in [5]:

$$l(\sigma, \xi; x) = \begin{cases} -n_u \log \sigma - \left(\frac{1}{\xi} + 1\right) \sum_{i=1}^{n_u} \log\left(1 + \frac{\xi x_i}{\sigma}\right), & \xi \neq 0 \\ \\ -n_u \log \sigma - \frac{1}{\sigma} \sum_{i=1}^{n_u} x_i, \qquad \xi = 0. \end{cases} \tag{9}$$

Analytical maximization of the log-likelihood for GPD is a challenge, numerical techniques are preferable to avoid numerical instabilities when $\xi = 0$. As shown by [3], in practice the maximum likelihood exists for $\xi \leq 1$.

After fitting the GPD to the exceedances, the parameters estimated $\hat{\xi}$ and $\hat{\sigma}$ are used to compute tail risk measures.

3.6 Dependence in Real-World Claim Data

A key assumption in this study is that insurance claim amounts are independent and identically distributed (i.i.d.), which is consistent with the asymptotic foundations of the POT method. However, in practice, claim data may exhibit temporal or cross-sectional dependence due to catastrophic events, seasonal effects, or policyholder-level clustering. Ignoring such dependence can lead to underestimation of tail risk and biased threshold selection.

Several approaches have been developed to address dependence in extremes. For instance, declustering techniques [5] identify approximately independent exceedances by grouping temporally close extremes. More recent work incorporates dependence structures directly, using copula-based extreme value models [1] or time-series extensions of EVT (e.g., extremal index methods). While these approaches provide a richer description of dependence, they often require more granular data and introduce additional modeling complexity.

Future research could extend the proposed logistic regression framework to account for dependence, for example, by integrating declustering steps prior to model fitting, or by incorporating covariates that

capture temporal or spatial effects. This would improve the robustness of threshold selection and risk estimation in operational insurance settings.

3.7 Model Validation

To assess the adequacy of the fitted GPD, we employed standard diagnostic plots. These include quantile–quantile (QQ) plots and probability plots, which compare the empirical distribution of the exceedances to the theoretical GPD. A good visual alignment between the empirical and theoretical quantiles indicates that the GPD provides an appropriate fit to the tail data.

3.8 Value at Risk and Expected Shortfall

Two risk measures commonly used in financial and actuarial practice are Value at Risk (VaR) and Expected Shortfall (ES). Both are derived from the fitted GPD and are used to quantify extreme losses at a given confidence interval α. These estimates follow the tail estimator described by [13].

Let n be the total number of observations and N_u be the number of exceedances above the threshold u. The tail probability function is:

$$\hat{F}(x) = 1 - \frac{N_u}{n}\left(1 + \frac{\xi(x-u)}{\sigma}\right)^{-1/\xi}. \tag{10}$$

The Value at Risk at the confidence level α, is given as:

$$\mathrm{VaR}_\alpha = u + \frac{\sigma}{\xi}\left(\left(\frac{n}{N_u}(1-\alpha)\right)^{-\xi} - 1\right). \tag{11}$$

The Expected Shortfall is given as:

$$\mathrm{ES}_\alpha = \frac{\mathrm{VaR}_\alpha}{1-\xi} + \frac{\sigma - \xi u}{1-\xi}. \tag{12}$$

In the context of health insurance, these risk metrics offer valuable information for assessing the financial impact of catastrophic medical events. For example, VaR_α at a high confidence level (e.g., 99%) helps estimate the claim amount that is unlikely to be exceeded, while ES_α provides the expected cost if that threshold is breached. These values support decision-making in premium setting, reserve allocation, and stop-loss policy design.

These risk measures are computed using the estimated GPD parameters and provide insights into potential losses in the tail of the claim distribution.

3.9 Algorithm: Threshold Selection Via Logistic Regression

The following steps summarize the proposed threshold selection approach:

1. **Initial threshold and labeling:** Define an initial threshold u_{init} as the 90th percentile of the observed claims. Label each claim X_i as defined in (5).
2. **Logistic model fitting:** Fit a logistic regression model to predict the probability of a claim being extreme as in (6).
3. **Probability based threshold selection:** For a pre-specified cutoff τ (e.g., 0.9), define the optimal threshold u^* as:
$$u^* = \min \{X_i : \hat{\pi}(X_i) \geq \tau\}.$$
4. **Exceedance Extraction and GPD Fitting:** Let $Z_i = X_i - u^*$ for all $X_i > u^*$. Fit a Generalized Pareto Distribution (GPD) to the exceedances using Maximum Likelihood Estimation.
5. **Tailrisk estimation:** Using the fitted GPD, compute VaR and ES at a 99% confidence level.

4 Results

The data set used for this study was obtained from Kaggle's open repository and contains anonymous individual health insurance records. It includes demographic and behavioral variables such as age, sex, body mass index (BMI), smoking status, region, and number of children, along with the total charges incurred by each policyholder. The charges variable represents the total medical expenses billed and serves as a suitable target for extreme value analysis due to its right-skewed distribution and the presence of extreme values. We assume that the claim amounts are independent and identically distributed, consistent with the theoretical foundations of the POT method in Extreme Value Theory [5, 11]. This assumption ensures the asymptotic validity of the Generalized Pareto Distribution for threshold exceedances.

The data set comprises 1,388 observations. The medical expenses range from approximately \$1,122 to over \$63,000. Figure 1 represents the distribution of total charges. The majority of individuals incur moderate costs, while a small number experience exceptionally high medical charges. This skewed distribution highlights the importance of accurately modeling the tail behavior, as extreme charges can impose significant financial risks to insurers.

Based on the nature of the claim values, choosing a suitable threshold is essential for modeling extremes. The Mean Residual Life (MRL) plot and a machine learning-based classification model using logistic regression were used.

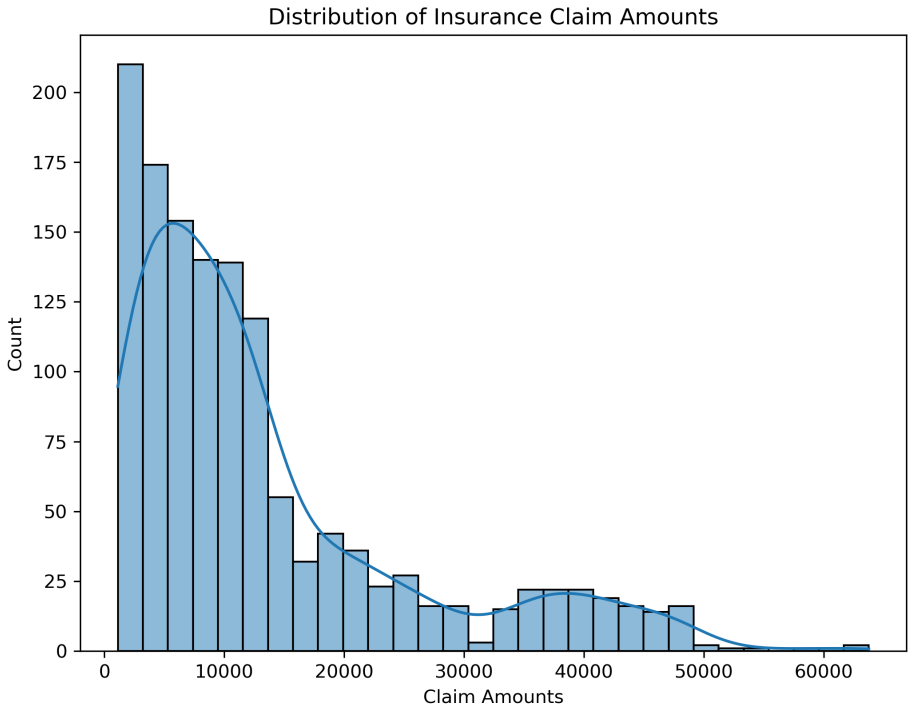

Figure 1: Distribution of insurance claim amounts.

4.1 Threshold Selection

To identify the optimal threshold, we first examined the MRL plot for regions of linearity. As shown in Figure 2, the MRL plot exhibits a linear region between $30,000$ and $40,000$. This suggests that the optimal threshold is within this range, and claim values exceeding a threshold within this range are well approximated using the generalized Pareto distribution.

While the MRL plot provides a visual guide for threshold selection, it does not necessarily state the actual optimal thresholds. Instead, the choice of an optimal threshold is often left to user discretion. We employed the machine learning approach to determine an optimal threshold objectively.

We defined the final threshold as the smallest claim with a predicted extremeness probability greater than 90%. This rule ensures a data-driven, interpretable, and reproducible method for threshold selection. The split and threshold selection procedure aligns with standard machine learning practices for detecting rare events.

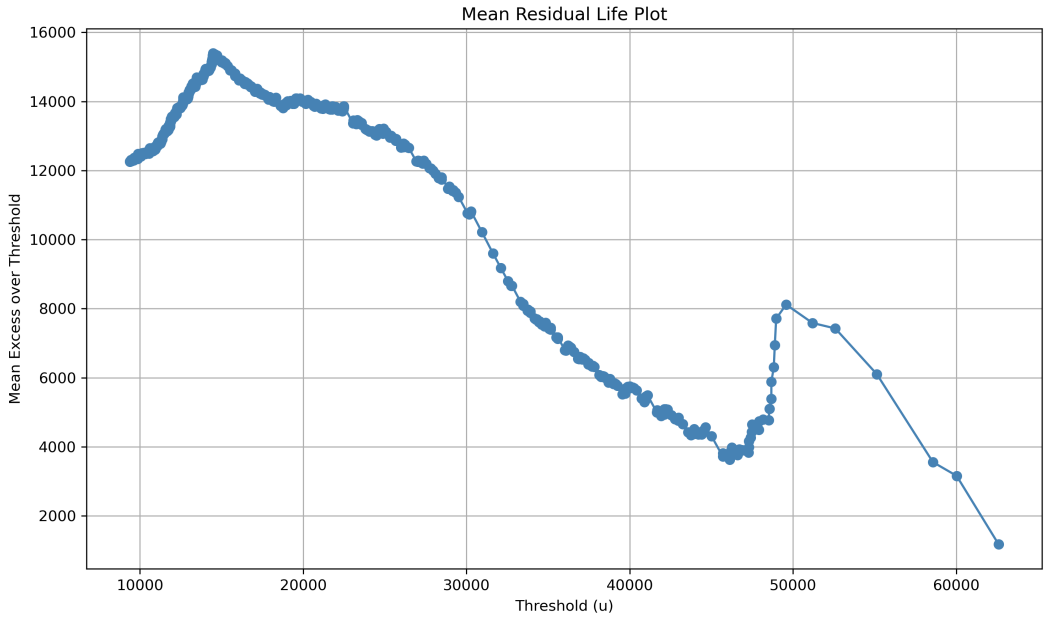

Figure 2: Mean Residual Life plot for insurance claim amounts.

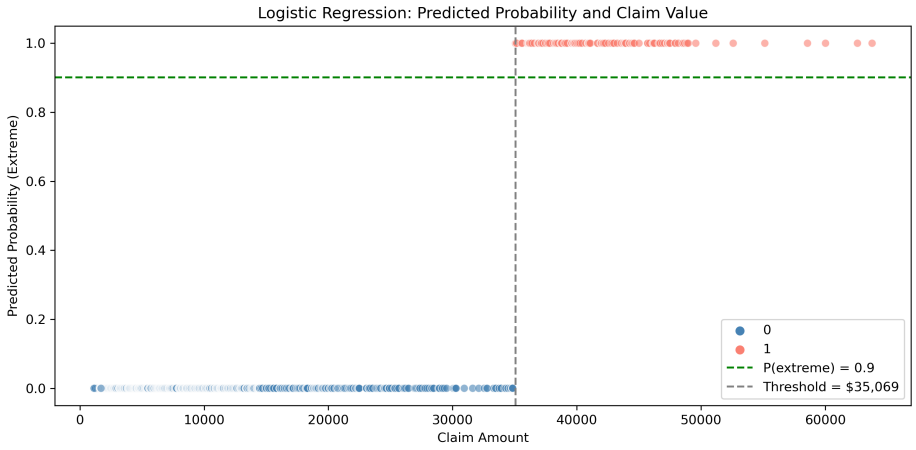

Figure 3: Predicted extremeness probabilities using logistic regression.

Figure 3 displays the predicted probability of the output of the scores by the logistic classifier. A claim is classified as extreme if the predicted probability is $\geq 90\%$. The horizontal green dashed lines indicate the probability threshold.

We observed that the model assigns a high probability near 1.0 to claims above a specific value. The vertical grey dotted line at 35, 069 is the minimum claim amount at which the model assigns a

predicted probability greater than 90%. The red dots represent claims classified as extreme (that is, with probability > 0.9), while the blue dots represent claims classified as nonextreme. This value, $u = 35,069$, selected via the logistic regression method, is used for subsequent modeling. This threshold aligns with the linear region previously identified in the MRL plot (Figure 2), confirming the insights from the graphical approach.

4.2 Sensitivity Analysis

To assess the robustness of the optimal threshold, we repeat the procedure using initial labeling quantiles 80%, 85%, 95%, and 99% . Table 1 shows the ML thresholds (u_{ML}), estimates of the GPD parameter, and tail risk measures 99%. The results indicated that the thresholds at the 80th and 85th percentiles were too low, leading to bias and a violation of the asymptotic assumption of the GPD. Moreso, thresholds at the 90th and 99th percentiles are too high, producing few exceedances and inflating variance. Based on this, we adopt the 90th percentile threshold as the final cut-off for subsequent modeling.

Table 1: Sensitivity of ML threshold and tail risk estimates across initial labeling quantiles.

q	u_{ML}	ξ	σ	$\text{VaR}_{0.99}$	$\text{ES}_{0.99}$
0.80	20,277.81	−0.4352	19,627.26	53,111.53	56,830.47
0.85	25,121.87	−0.0031	12,168.16	56,752.34	68,711.55
0.95	41,661.60	−0.0727	5,343.07	49,975.10	54,140.38
0.99	48,549.18	1.9171	645.40	48,530.18	∞

4.3 Fitting Generalized Pareto Model

After selecting an optimal threshold, we extracted the exceedances, defined as

$$y_i = X_i - u, \quad \text{for all } X_i > u.$$

The generalized Pareto model was fitted to the exceedances of the claim values and the parameters were estimated using the maximum likelihood estimation. The estimated values obtained are:

$$\hat{\xi} = -0.25, \quad \hat{\sigma} = 9,124.73.$$

The negative shape parameter ($\hat{\xi} < 0$) indicates a short-tailed distribution, suggesting the distribution has an upper bound. This is consistent with real-world expectations for insurance claims, which typically have natural policy limits. To assess the model adequacy, we evaluated the fit using the QQ plot and the PP plot shown in Figure 4. Most of the points in both plots lie within the 90% confidence interval, indicating that the GPD provides a good fit to the exceedances based on the selected threshold.

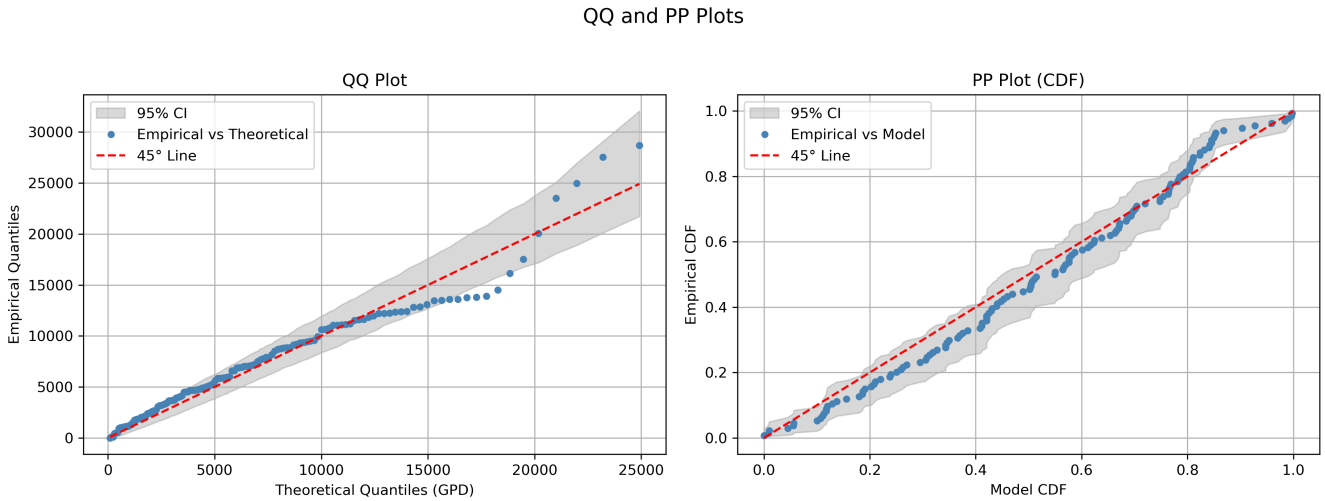

Figure 4: Diagnostic plots.

4.4 Estimate of VaR and ES

To quantify the tail risk of extreme insurance claims, we estimate the VaR and ES at the 99% confidence level using the fitted GPD model. The VaR represents the minimum loss that can be expected with 99% confidence interval. While the ES captures the expected severity of losses that exceed the VaR threshold. The estimates obtained are:

$$\text{VaR} = \$51,291.44, \quad \text{ES} = \$65,728.25.$$

This indicates that 99% of future insurance claims are expected to fall below $\$51,291.44$, however in rare cases, the average claim is expected to exceed $\$65,728.25$.

5 Conclusion and Recommendation

This study introduced a machine learning-assisted threshold selection strategy to enhance the application of Extreme Value Theory. Unlike traditional approaches, such as the MRL plot, which require visual interpretation and may vary between practitioners, the proposed logistic regression-based method uses a probability-based rule to determine the threshold. This enables full reproducibility, consistency across users, and ease of integration into automated actuarial workflows.

Using insurance claim data for demonstration, we showed that the selected threshold aligned with the linear region of the MRL plot, thereby validating the approach. Based on this threshold, we estimated the Value at Risk (VaR) and Expected Shortfall (ES) at the 99% confidence level to quantify the tail risk of

extreme insurance claims. This framework offers a robust and interpretable alternative for tail modeling and can be adapted to other domains.

A primary limitation of this study is the assumption that insurance claim amounts are independent and identically distributed as required by classical EVT. While reasonable for static claim datasets, this assumption may not hold in real-world situations involving temporal clustering, policy changes, or catastrophic events. In such cases, ignoring dependence structures could lead to biased threshold selection and inaccurate tail estimates.

Future research should consider extending this framework to accommodate dependent data structures. One practical recommendation is to collect data with sufficient temporal granularity to assess and address dependence through declustering methods. For instance, the use of run declustering or block maxima with temporal buffers may help isolate independent exceedances, thereby preserving the theoretical underpinnings of EVT while adapting to real-world complexities.

Moreover, integrating machine learning models that account for temporal or spatial correlations would make the approach more robust in operational environments such as reinsurance pricing, catastrophe modeling, and systemic risk assessment.

References

[1] Beirlant, J., Goegebeur, Y., Segers, J., & Teugels, J. L. (2006). *Statistics of extremes: Theory and applications*. John Wiley & Sons. https://doi.org/10.1002/0470012382

[2] Bommier, E. (2014). Peaks-over-threshold modelling of environmental data (Technical report, U.U.D.M. Project Report 2014:33). Department of Mathematics, Uppsala University.

[3] del Castillo, J., & Daoudi, J. (2009). Estimation of the generalized Pareto distribution. *Statistics & Probability Letters, 79*(5), 684–688. https://doi.org/10.1016/j.spl.2008.10.021

[4] Charras-Garrido, M., & Lezaud, P. (2013). Extreme value analysis: An introduction. *Journal de la Société Française de Statistique, 154*(2), 66–97.

[5] Coles, S., Bawa, J., Trenner, L., & Dorazio, P. (2001). *An introduction to statistical modeling of extreme values* (Vol. 208). Springer. https://doi.org/10.1007/978-1-4471-3675-0

[6] Davison, A. C., & Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 52*(3), 393–425. https://doi.org/10.1111/j.2517-6161.1990.tb01796.x

[7] Embrechts, P., Klüppelberg, C., & Mikosch, T. (1997). *Modelling extremal events: For insurance and finance* (Vol. 33). Springer. https://doi.org/10.1007/978-3-642-33483-2

[8] Hosmer, D. W. Jr., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). John Wiley & Sons. https://doi.org/10.1002/9781118548387

[9] Mosala, R., Rachuene, K. A., & Shongwe, S. C. (2024). Most suitable threshold method for extremes in financial data with different volatility levels. *ITM Web of Conferences*, *67*, 01033. EDP Sciences. https://doi.org/10.1051/itmconf/20246701033

[10] Northrop, P. J., & Coleman, C. L. (2014). Improved threshold diagnostic plots for extreme value analyses. *Extremes, 17*, 289–303. https://doi.org/10.1007/s10687-014-0183-z

[11] Pickands, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics, 3*(1), 119–131. https://doi.org/10.1214/aos/1176343003

[12] Scarrott, C., & MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT–Statistical Journal, 10*(1), 33–60. https://doi.org/10.57805/revstat.v10i1.110

[13] Singh, A. K., Allen, D. E., & Powell, R. J. (2011). Value at Risk estimation using Extreme Value Theory. In F. Chan, D. Marinova, & R. S. Anderssen (Eds.), MODSIM 2011: 19th International Congress on Modelling and Simulation: Proceedings (pp. 1478â1484). Modelling and Simulation Society of Australia and New Zealand.

[14] Tang, Y., Wang, H. J., & Li, D. (2024). High-dimensional extreme quantile regression. *arXiv preprint arXiv:2411.13822*. Retrieved from https://arxiv.org/abs/2411.13822

[15] Wadsworth, J. L. (2016). Exploiting structure of maximum likelihood estimators for extreme value threshold selection. *Technometrics, 58*(1), 116–126. https://doi.org/10.1080/00401706.2014.998345

[16] Wager, S., & Athey, S. (2024). Extremal random forests. *Journal of the American Statistical Association, 119*(548), 1–24. https://doi.org/10.1080/01621459.2023.2300522

[17] World Meteorological Organization. (2009). Guidelines on analysis of extremes in a changing climate in support of informed decisions for adaptation (WMO-No. 1009).

This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted, use, distribution and reproduction in any medium, or format for any purpose, even commercially provided the work is properly cited.