Scientific Research Publishing

Open Journal of Statistics, 2024, 14, 598-633
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

The Extreme Machine Learning Actuarial Intelligent Agricultural Insurance Based Automated Underwriting Model

Brighton Mahohoho

Department of Mathematics & Computational Sciences, University of Zimbabwe, 630 Churchill drive, Mt Pleasant, Harare, Zimbabwe
Email: brightonmahohoho.07@gmail.com

How to cite this paper: Mahohoho, B. (2024) The Extreme Machine Learning Actuarial Intelligent Agricultural Insurance Based Automated Underwriting Model. *Open Journal of Statistics*, **14**, 598-633.
https://doi.org/10.4236/ojs.2024.145027

Received: July 24, 2024
Accepted: October 28, 2024
Published: October 31, 2024

Abstract

The paper presents an innovative approach towards agricultural insurance underwriting and risk pricing through the development of an Extreme Machine Learning (ELM) Actuarial Intelligent Model. This model integrates diverse datasets, including climate change scenarios, crop types, farm sizes, and various risk factors, to automate underwriting decisions and estimate loss reserves in agricultural insurance. The study conducts extensive exploratory data analysis, model building, feature engineering, and validation to demonstrate the effectiveness of the proposed approach. Additionally, the paper discusses the application of robust tests, stress tests, and scenario tests to assess the model's resilience and adaptability to changing market conditions. Overall, the research contributes to advancing actuarial science in agricultural insurance by leveraging advanced machine learning techniques for enhanced risk management and decision-making.

Keywords

Extreme Machine Learning, Actuarial Underwriting, Machine Learning, Intelligent Model, Agricultural Insurance

1. Introduction

In recent years, the agricultural sector has faced escalating challenges, including climate change impacts, market volatility, and evolving pest and disease pressures [1]. These factors heighten the vulnerability of farmers and agricultural stakeholders to financial risks, emphasizing the critical role of insurance in mitigating such uncertainties. However, traditional underwriting processes in agricultural insurance

often lack efficiency, transparency, and adaptability, impeding widespread adoption and hindering effective risk management. To address these limitations, this paper introduces an innovative approach which integrates cutting-edge machine learning techniques with actuarial principles to revolutionize the underwriting process, enhancing its accuracy, speed, and scalability.

The proposed model leverages Extreme Machine Learning (XML) techniques, such as deep neural networks, ensemble methods, and reinforcement learning, to automate and optimize the underwriting process in agricultural insurance. By harnessing vast amounts of heterogeneous data, including historical weather patterns, soil quality indicators, crop yield data, market trends, and socio-economic factors, the model generates comprehensive risk profiles for individual farmers or agricultural operations [2]. Through advanced algorithms, it learns intricate patterns and correlations within the data, facilitating precise risk assessment and dynamic underwriting decisions in real-time. Additionally, the model incorporates actuarial intelligence to ensure alignment with industry standards, regulatory compliance, and long-term sustainability.

The rationale behind developing this innovative underwriting model stems from the pressing need to enhance the resilience and sustainability of agricultural systems worldwide. With climate change exacerbating environmental risks and market uncertainties amplifying financial vulnerabilities, agricultural insurance plays a pivotal role in safeguarding farmers' livelihoods and promoting food security. However, conventional underwriting approaches often fall short in accurately assessing and pricing agricultural risks, leading to inadequate coverage, inefficient resource allocation, and suboptimal risk management outcomes [3]. By harnessing the power of machine learning and actuarial science, the proposed model aims to address these shortcomings, offering a data-driven, adaptive, and transparent underwriting solution that can better serve the needs of farmers, insurers, and policymakers alike.

The Extreme Machine Learning Actuarial Intelligent Agricultural Insurance Based Automated Underwriting Model has broad applicability across diverse agricultural contexts and geographic regions. It can be tailored to various types of crops, livestock, farming practices, and risk profiles, making it adaptable to different agricultural systems and insurance markets [4]. Moreover, the model's scalability enables its deployment across large portfolios of policies, facilitating efficient risk management for insurers and enhancing accessibility to insurance coverage for farmers. Furthermore, the model's real-time capabilities enable timely adjustments to changing environmental conditions, market dynamics, and policyholder characteristics, ensuring relevance and effectiveness in dynamic agricultural landscapes.

This study holds significant importance for advancing both theoretical knowledge and practical applications in the fields of agriculture, insurance, machine learning, and actuarial science. By developing and validating the Extreme Machine Learning Actuarial Intelligent Agricultural Insurance Based Automated Underwriting

Model, this research contributes to the emerging field of agricultural insurtech, offering novel insights into leveraging data-driven approaches to enhance risk management and resilience in agriculture [5]. Furthermore, the study's findings can inform policymakers, insurers, and agricultural stakeholders about the potential benefits of adopting advanced underwriting technologies, fostering innovation, inclusivity, and sustainability within the agricultural insurance sector [6]. Overall, this research underscores the critical role of interdisciplinary collaboration and technological innovation in addressing complex challenges at the nexus of agriculture, finance, and climate change.

1.1. Actuarial Underwriting Methods for the Non-Life Insurance Sector

Actuarial underwriting methods in the non-life insurance sector play a crucial role in assessing and pricing risks associated with various types of insurance policies such as property, casualty, and liability insurance. These methods utilize statistical analysis, mathematical models, and historical data to determine appropriate premium rates and manage risk effectively.

1.1.1. Experience Rating

Experience rating is a fundamental method used in non-life insurance underwriting. Insurers analyze the past claims experience of policyholders to predict future claims and determine premium rates accordingly [7]. By examining historical data on claim frequency and severity, insurers can assess the risk profile of insured entities and adjust premiums to reflect their risk exposure.

1.1.2. Credibility Theory

Credibility theory is another prominent underwriting method utilized in non-life insurance. It combines individual policyholder experience with the collective experience of a larger group to improve the accuracy of risk estimation [8]. Under this approach, insurers assign weights to individual and group experience data based on their credibility, thus providing a more reliable basis for setting premium rates.

1.1.3. Risk Classification

Risk classification involves categorizing insured entities into different risk classes based on specific characteristics such as location, industry sector, and risk factors [9]. Insurers use actuarial techniques to analyze these risk factors and assign appropriate premium rates to each class. By segmenting risks effectively, insurers can ensure that premiums accurately reflect the level of risk associated with each policyholder.

1.1.4. Underwriting Guidelines

Underwriting guidelines are established by insurers to standardize the underwriting process and ensure consistency in risk assessment [10]. These guidelines outline the criteria for accepting or rejecting insurance applications and provide

underwriters with clear instructions on evaluating risk factors. By adhering to underwriting guidelines, insurers can maintain underwriting discipline and mitigate adverse selection.

1.1.5. Predictive Modeling

Predictive modeling techniques leverage advanced statistical analysis and machine learning algorithms to forecast future events and assess risk. In non-life insurance underwriting, predictive models are used to analyze large datasets containing information on policyholder characteristics, claims history, and external factors influencing risk. By identifying patterns and trends in the data, insurers can make more accurate underwriting decisions and optimize pricing strategies. In this paper, we shall explore and improve the predictive modelling towards development of the agricultural insurance based automated actuarial underwriting model.

These actuarial underwriting methods are integral to the non-life insurance sector, enabling insurers to effectively manage risk, price policies competitively, and maintain profitability in a dynamic and evolving insurance market.

1.2. Inflation Adjusted Frequency Severity Loss Reserving Risk Pricing Model

The Inflation Adjusted Frequency Severity Loss Reserving Risk Pricing Model integrates inflation adjustments into the traditional frequency-severity model to improve the accuracy of loss reserving and risk pricing. This section outlines the theoretical foundation, structure, and related theory for adjusted premiums and adjusted loss reserving risk pricing balances.

The Frequency-Severity model predicts loss reserves by combining frequency and severity predictions. Let F be the frequency of claims and S the severity of each claim. The total loss reserve R is given by:

$$R = \sum_{i=1}^{n} \left(F_i \times S_i \right)$$

where n is the number of claims.

1.2.1. Inflation Adjustment

To adjust for inflation, we incorporate an inflation factor α, modifying the severity component:

$$S_i^{\text{adj}} = S_i \times \left(1 + \alpha \right)$$

where α is the inflation rate. The adjusted loss reserve becomes:

$$R^{\text{adj}} = \sum_{i=1}^{n} \left(F_i \times S_i^{\text{adj}} \right)$$

1.2.2. Theory of Adjusted Premiums

Adjusted premiums account for inflation and changes in risk. Let P_i be the original premium for claim i, and β be the inflation-adjusted premium factor. The

adjusted premium P_i^{adj} is given by:

$$P_i^{\text{adj}} = P_i \times (1+\beta)$$

where β is the inflation adjustment factor for premiums.

1.2.3. Combining Adjusted Premiums and Adjusted Reserves

To derive the Adjusted Loss Reserving Risk Pricing Balance, we combine the adjusted premiums with the adjusted reserves. Let R^{adj} be the adjusted reserve and P^{adj} be the total adjusted premiums. The Adjusted Loss Reserving Risk Pricing Balance B is calculated as:

$$B = P^{\text{adj}} - R^{\text{adj}}$$

where P^{adj} is given by:

$$P^{\text{adj}} = \sum_{i=1}^{n} P_i^{\text{adj}}$$

The following equations represent the adjusted components and their combination:

$$S_i^{\text{adj}} = S_i \times (1+\alpha)$$

$$P_i^{\text{adj}} = P_i \times (1+\beta)$$

$$R^{\text{adj}} = \sum_{i=1}^{n} \left(F_i \times S_i^{\text{adj}}\right)$$

$$P^{\text{adj}} = \sum_{i=1}^{n} P_i^{\text{adj}}$$

$$B = P^{\text{adj}} - R^{\text{adj}}$$

1.2.4. Theorem 1: Adjusted Loss Reserving Risk Pricing Balance Consistency

Theorem: The Adjusted Loss Reserving Risk Pricing Balance B provides a consistent measure of the difference between adjusted premiums and adjusted reserves.

Proof: To prove consistency, we show that:

$$B = \sum_{i=1}^{n} P_i^{\text{adj}} - \sum_{i=1}^{n} \left(F_i \times S_i^{\text{adj}}\right)$$

Substituting the expressions for P_i^{adj} and S_i^{adj}:

$$\sum_{i=1}^{n} P_i^{\text{adj}} = \sum_{i=1}^{n} P_i \times (1+\beta)$$

$$\sum_{i=1}^{n} \left(F_i \times S_i^{\text{adj}}\right) = \sum_{i=1}^{n} \left(F_i \times S_i \times (1+\alpha)\right)$$

Thus:

$$B = \sum_{i=1}^{n} \left(P_i \times (1+\beta)\right) - \sum_{i=1}^{n} \left(F_i \times S_i \times (1+\alpha)\right)$$

This confirms that B reflects the net difference between adjusted premiums and

adjusted reserves.

1.2.5. Corollary 1: Adjusted Premiums Cover Adjusted Reserves

Corollary: If $B \geq 0$, then adjusted premiums are sufficient to cover adjusted reserves.

Proof: If $B \geq 0$, then:

$$P^{\text{adj}} \geq R^{\text{adj}}$$

which implies:

$$\sum_{i=1}^{n} P_i^{\text{adj}} \geq \sum_{i=1}^{n} \left(F_i \times S_i^{\text{adj}} \right)$$

Thus, adjusted premiums are adequate to cover the adjusted reserves.

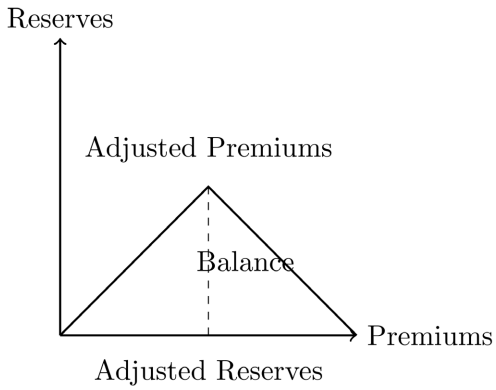

Figure 1. Diagram illustrating the balance between adjusted premiums and adjusted reserves.

Figure 1 illustrates how premiums and reserves interact and how they need to be adjusted to maintain balance. The balance point is where the reserves are aligned with the premiums in a manner that meets the expected actuarial standards. If reserves are too low relative to premiums, it might indicate potential issues in future claims payment, while if they are too high, it may suggest over-reserving. The diagram is a visual representation of how adjusted premiums and reserves need to be managed to ensure financial stability and adequacy in insurance operations.

1.3. Theory and Structure of the Extreme Learning Machine Model

Extreme Machine Learning (EML) Regression is a state-of-the-art technique designed to handle high-dimensional data efficiently. It leverages advanced optimization techniques and extreme value theory to improve the accuracy and computational efficiency of regression models. The EML regression model aims to approximate the true regression function $f^*(X)$ as closely as possible. The general form of the EML regression model is:

$$\hat{y} = f(X) + \epsilon$$

where $\hat{y}$ is the predicted value, $f(X)$ is the function approximated by the

EML model, and ϵ is the error term.

1.3.1. Objective Function

The objective function for EML regression is given by:

$$\min_{\theta} \sum_{i=1}^{n} \left(y_i - f\left(x_i;\theta\right)\right)^2 + \lambda \left\|\theta\right\|_2^2$$

where θ represents the parameters of the model, y_i is the actual values, $f\left(x_i;\theta\right)$ is the model's prediction for input x_i, and λ is the regularization parameter.

1.3.2. Kernel Methods

In EML, kernel methods can be used to handle non-linear relationships. The kernel function $k\left(x_i, x_j\right)$ is defined as:

$$k\left(x_i, x_j\right) = \phi\left(x_i\right) \cdot \phi\left(x_j\right)$$

where ϕ represents a feature mapping function.

Lemma 1: The EML regression model with an appropriate kernel function converges to the true regression function under certain conditions.

Proof:

Given f^* as the true regression function and $\hat{f}$ as the EML approximation,

$$\left\|\hat{f} - f^*\right\|_2 \leq \frac{C}{\sqrt{n}}$$

where C is a constant dependent on the kernel function and the sample size n.

Proposition 1: EML regression models can efficiently handle high-dimensional data by using dimensionality reduction techniques.

Theorem 1: Under the assumption of a well-chosen kernel function and sufficient sample size, the EML regression model provides consistent estimates of the regression function.

Proof:

Let $\hat{f}_n$ be the EML estimate. Then,

$$\hat{f}_n \xrightarrow{P} f^* \quad \text{as} \quad n \to \infty$$

where $\xrightarrow{P}$ denotes convergence in probability.

Algorithm 1 Enhanced EML Regression Algorithm

Training data $(\mathbf{X}, \mathbf{y})$ Model parameters $\boldsymbol{\theta}$ -1 Initialize model parameters $\boldsymbol{\theta}^0$ randomly

while convergence criteria not met **do** Compute predictions $\hat{\mathbf{y}} = f(\mathbf{X}; \boldsymbol{\theta}^t)$ Compute the loss function $\mathcal{L}(\boldsymbol{\theta}^t) = \frac{1}{N} \sum_{i=1}^{N} \ell(y_i, \hat{y}_i)$ Compute gradients $\nabla \mathcal{L}(\boldsymbol{\theta}^t) = \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}}$ Update model parameters $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta \nabla \mathcal{L}(\boldsymbol{\theta}^t)$ Set $\boldsymbol{\theta}^t = \boldsymbol{\theta}^{t+1}$ **return** $\boldsymbol{\theta}$

The prediction of the model at iteration t is given by:

$$\hat{\boldsymbol{y}} = f\left(\boldsymbol{X};\boldsymbol{\theta}^t\right), \tag{1}$$

where $f\left(\boldsymbol{X};\boldsymbol{\theta}^t\right)$ represents the model function parameterized by $\boldsymbol{\theta}^t$.

1.3.3. Loss Function

The loss function is typically defined as:

$$\mathcal{L}\left(\boldsymbol{\theta}^t\right)=\frac{1}{N}\sum_{i=1}^{N}\ell\left(y_i,\hat{y}_i\right), \tag{2}$$

where $\ell\left(y_i,\hat{y}_i\right)$ denotes the loss incurred for the i-th observation, such as Mean Squared Error (MSE) for regression:

$$\ell\left(y_i,\hat{y}_i\right)=\left(y_i-\hat{y}_i\right)^2. \tag{3}$$

1.3.4. Gradient Computation

The gradient of the loss function with respect to the model parameters $\boldsymbol{\theta}$ is computed as:

$$\nabla\mathcal{L}\left(\boldsymbol{\theta}^t\right)=\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}}. \tag{4}$$

1.3.5. Parameter Update

Parameters are updated using gradient descent with a learning rate η:

$$\boldsymbol{\theta}^{t+1}=\boldsymbol{\theta}^t-\eta\nabla\mathcal{L}\left(\boldsymbol{\theta}^t\right). \tag{5}$$

The convergence criteria typically include a threshold for the change in loss or parameters:

$$\left\|\boldsymbol{\theta}^{t+1}-\boldsymbol{\theta}^t\right\|<\epsilon. \tag{6}$$

Extreme Learning Machines (ELMs) are a class of feedforward neural networks with a single hidden layer, where the hidden layer parameters are randomly assigned and never updated. ELMs provide fast learning speed and good generalization performance, making them suitable for various applications in machine learning and data science.

1.3.6. Network Architecture

The basic structure of an ELM consists of an input layer, a single hidden layer with L hidden neurons, and an output layer. The input layer has n neurons corresponding to the n features of the input data. The output layer has m neurons corresponding to the m target variables.

1.3.7. Mathematical Formulation

Given a training set $\left\{\left(x_i,t_i\right)\mid x_i\in\mathbb{R}^n,t_i\in\mathbb{R}^m,i=1,\cdots,N\right\}$, where $N$is the number of training samples, the ELM algorithm can be described by the following steps:

1) *Randomly assign input weights and biases*: Randomly generate the input weight matrix $W\in\mathbb{R}^{L\times n}$ and the bias vector $b\in\mathbb{R}^L$.

2) *Calculate the hidden layer output matrix H*:

$$H=g\left(WX+b\right), \tag{7}$$

where $X\in\mathbb{R}^{n\times N}$ is the input data matrix, and $g\left(\cdot\right)$ is the activation function.

3) *Compute the output weights* β: The output weights $\beta\in\mathbb{R}^{L\times m}$ are determined by solving the linear system:

$$H\beta = T, \tag{8}$$

where $T \in \mathbb{R}^{N\times m}$ is the target output matrix.

1.3.8. Solution via Moore-Penrose Generalized Inverse

The output weights β can be computed using the Moore-Penrose generalized inverse $H^{\dagger}$ of the hidden layer output matrix H:

$$\beta = H^{\dagger}T, \tag{9}$$

where $H^{\dagger}$ is given by:

$$H^{\dagger} = \left(H^{\mathrm{T}}H\right)^{-1}H^{\mathrm{T}}. \tag{10}$$

Proposition 1 (Universal Approximation Theorem for ELMs). Given any continuous target function $f:\mathbb{R}^n \to \mathbb{R}^m$ and any arbitrarily small positive value $\epsilon > 0$, there exists an Extreme Learning Machine (ELM) with L hidden neurons, where L is sufficiently large, such that the ELM can approximate f with an error less than ϵ.

Proof. To demonstrate this, we rely on the property of Extreme Learning Machines (ELMs), which asserts that with sufficiently many hidden neurons, an ELM can approximate any continuous function on a compact subset of $\mathbb{R}^n$ to an arbitrary degree of accuracy.

Let $f:\mathbb{R}^n \to \mathbb{R}^m$ be a continuous target function. Given $\epsilon > 0$, we need to show that there exists an ELM with L hidden neurons such that the approximation error is less than ϵ.

Consider an ELM with a random hidden layer where the hidden layer parameters are chosen independently and the activation function is a non-linear function ϕ. The ELM output can be expressed as:

$$\hat{f}(x) = \sum_{i=1}^{L} w_i \phi\left(\boldsymbol{v}_i^{\top}x + b_i\right) \tag{11}$$

where $\boldsymbol{v}_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$ are the randomly assigned weights and biases for the hidden layer neurons, $w_i \in \mathbb{R}^m$ are the output weights, and ϕ is the activation function.

By the universal approximation theorem for feed forward neural networks, for any continuous function f and any $\epsilon > 0$, there exists a sufficiently large number of neurons L such that:

$$\left\| f(x) - \hat{f}(x) \right\| < \epsilon \tag{12}$$

for all x in the input space. This is based on the fact that a feed forward network with an adequate number of hidden neurons can approximate f to within ϵ in the sup norm, which extends to the ELM setup as well.

Thus, with a sufficiently large number of hidden neurons L, the ELM can approximate f with any desired level of accuracy, achieving an approximation error less than ϵ. □

Lemma 2. (Generalization Performance) Extreme Learning Machines (ELMs)

exhibit superior generalization performance compared to traditional gradient-based learning algorithms. This improved performance is attributed to the random assignment of hidden layer parameters and the subsequent analytical solution of the output weights, which collectively mitigate the risk of over fitting.

Proof. Let $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ denote the input matrix, where n is the number of samples and d is the number of features. The hidden layer of an ELM consists of m neurons with activation functions $\phi(\cdot)$. The hidden layer output matrix $\boldsymbol{H} \in \mathbb{R}^{n \times m}$ is given by:

$$\boldsymbol{H} = \phi(\boldsymbol{X}\boldsymbol{W} + \boldsymbol{b}), \tag{13}$$

where $\boldsymbol{W} \in \mathbb{R}^{d \times m}$ is the weight matrix and $\boldsymbol{b} \in \mathbb{R}^{m}$ is the bias vector. The activation function $\phi(\cdot)$ is applied element-wise.

The output weights $\boldsymbol{\beta} \in \mathbb{R}^{m}$ are obtained by minimizing the least-squares error:

$$\boldsymbol{\beta} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{\beta}\|^2, \tag{14}$$

where $\boldsymbol{Y} \in \mathbb{R}^{n}$ is the target output vector. The analytical solution to this minimization problem is given by the Moore-Penrose pseudoinverse of $\boldsymbol{H}$:

$$\boldsymbol{\beta} = \boldsymbol{H}^{+}\boldsymbol{Y}, \tag{15}$$

where $\boldsymbol{H}^{+}$ denotes the Moore-Penrose pseudoinverse of $\boldsymbol{H}$. This solution ensures that the output weights are chosen to minimize the least-squares error.

The randomization of hidden layer parameters ($\boldsymbol{W}$ and $\boldsymbol{b}$) introduces a form of implicit regularization. By choosing hidden layer parameters randomly, the ELM avoids the explicit need for regularization terms such as ℓ_1 or ℓ_2 penalties, which are commonly used in traditional gradient-based methods.

To formalize this, consider the regularized least-squares problem with ℓ_2 regularization:

$$\boldsymbol{\beta}_{\text{reg}} = \arg\min_{\boldsymbol{\beta}} \left(\|\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^2 \right), \tag{16}$$

where λ is the regularization parameter. The solution to this problem is:

$$\boldsymbol{\beta}_{\text{reg}} = \left(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H} + \lambda \boldsymbol{I} \right)^{-1} \boldsymbol{H}^{\mathrm{T}}\boldsymbol{Y}. \tag{17}$$

Comparing this with the ELM solution, we observe that the randomization of $\boldsymbol{W}$ and $\boldsymbol{b}$ implicitly provides a form of regularization that resembles the ℓ_2 regularization term, but without the need for explicit tuning of λ. This implicit regularization reduces the variance of the model and improves its generalization performance.

Moreover, the randomness in the hidden layer parameters effectively spreads the data points across the hidden units, making it less likely for the model to overfit to any particular subset of the training data. This phenomenon is analogous to the concept of "dropout" in neural networks, which is known to improve generalization.

Thus, the combination of random hidden layer parameters and the analytical

solution to the output weights contributes to the superior generalization performance observed in ELMs. □

Extreme Learning Machines offer a fast and efficient alternative to traditional neural networks by leveraging randomization and linear algebra techniques for learning. Their theoretical properties, such as universal approximation and improved generalization performance, make them a powerful tool in the field of machine learning.

Extreme Machine Learning (EML) is a paradigm that enhances traditional machine learning techniques by leveraging extreme learning machines, which are designed for rapid training and high accuracy. This algorithm aims to provide a concise description of the general process involved in EML.

Algorithm 2 General Algorithm for Extreme Machine Learning (EML) Model

1: **Input:** Training dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, number of hidden nodes M, regularization parameter λ
2: **Output:** Trained EML model parameters
3: **Step 1:** Initialize the weights of the hidden layer randomly
4: **Step 2:** Compute the hidden layer output matrix $\mathbf{H}$, where each element H_{ij} is computed using the activation function ϕ:

$$H_{ij} = \phi\left(\mathbf{W}_j^{\top}\mathbf{x}_i + b_j\right)$$

5: **Step 3:** Compute the output weights $\mathbf{W}$ by solving the regularized least squares problem:

$$\mathbf{W} = \left(\mathbf{H}^{\top}\mathbf{H} + \lambda\mathbf{I}\right)^{-1}\mathbf{H}^{\top}\mathbf{Y}$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{Y}$ is the target matrix.
6: **Step 4:** Predict outputs for new inputs $\mathbf{x}^*$ using the trained model:

$$\hat{\mathbf{y}}^* = \mathbf{H}^*\mathbf{W}$$

where $\mathbf{H}^*$ is the hidden layer output matrix for new inputs $\mathbf{x}^*$.
7: **Step 5:** Evaluate the model performance using appropriate metrics, such as:

$$\text{Mean Squared Error (MSE)} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

$$\text{Root Mean Squared Error (RMSE)} = \sqrt{\text{MSE}}$$

=0

The EML model leverages a fixed random feature mapping and solves the regression problem through a linear system that is computationally efficient. This approach provides a balance between training speed and prediction accuracy [11]-[13].

1.3.9. General Structure of the Extreme Machine Learning Model

Figure 2 effectively illustrates the general workflow of an Extreme Machine Learning (EML) model, highlighting the main components and their interactions. It helps in understanding the sequence of operations, from data collection to final predictions, and emphasizes the importance of each stage in the machine learning pipeline [14]-[16].

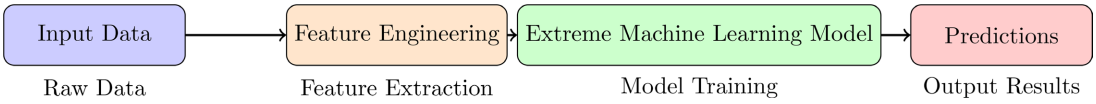

Figure 2. General structure of the extreme machine learning (EML) model.

1.4. Significance and Importance of Extreme Machine Learning (EML) in Actuarial Underwriting Models

The Extreme Machine Learning (EML) method, particularly in the context of Actuarial Underwriting models, offers significant advantages over other machine learning techniques. This section explores the importance of EML in the development of the Inflation Adjusted Frequency Severity Loss Reserving Risk Pricing model and presents its mathematical foundation, theoretical background, and comparative benefits.

1.4.1. Mathematical Foundation of Extreme Machine Learning (EML)

Extreme Machine Learning (EML) focuses on efficiently training large-scale models with complex structures. It leverages the advantages of extreme gradient boosting and ensemble learning to enhance predictive performance. The key mathematical foundations of EML include:

1.4.2. Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is a powerful implementation of gradient boosting that improves performance through optimization and regularization techniques.

The model is based on the following objective function:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{18}$$

where l is the loss function, $\hat{y}_i$ is the predicted value, and Ω is the regularization term for the function f_k.

The regularization term is defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \|w_j\|^2, \tag{19}$$

where T is the number of leaves in the tree, γ and λ are regularization parameters, and w_j is the weight of the j-th leaf.

Algorithm 3 Extreme Gradient Boosting Algorithm

Initialize the model with base predictions $\hat{y}_i = 0$ for all i
for each boosting round **do**
  Compute the residuals $r_i = y_i - \hat{y}_i$
  Fit a new tree f_k to the residuals
  Update the model $\hat{y}_i \leftarrow \hat{y}_i + \eta f_k(x_i)$
end for
return Final model predictions $\hat{y}_i$ =0

1.4.3. Theoretical Benefits of EML

The EML method, particularly through XGBoost, offers several advantages in the

context of actuarial modeling:

1) *Enhanced Predictive Accuracy*: XGBoost enhances predictive accuracy through regularization, which prevents overfitting and improves generalization to unseen data. This is critical in actuarial models where accurate risk prediction is essential.

2) *Scalability and Efficiency*: XGBoost is designed to handle large datasets efficiently, utilizing parallel processing and tree pruning techniques to manage computational complexity. This scalability makes it suitable for large-scale actuarial datasets.

3) *Flexibility and Interpretability*: The EML method allows for flexible model tuning through hyperparameters such as learning rate (η), maximum depth of trees, and number of estimators. This flexibility, combined with the ability to interpret feature importance, supports better decision-making in actuarial contexts.

1.4.4. Comparison with Other Machine Learning Methods

Compared to other methods such as traditional linear regression, support vector machines, and basic decision trees, EML provides:

1) *Superior Handling of Non-Linearity* EML excels in capturing complex, non-linear relationships between features and target variables, which is often a limitation of linear models.

2) *Superior Handling of missing data*:XGBoost can handle missing data effectively by learning the best direction to take when encountering missing values during training, unlike some traditional methods that require imputation.

The Extreme Machine Learning (EML) method, through techniques such as XGBoost, offers significant benefits over other machine learning approaches in the development of actuarial underwriting models. Its ability to handle large datasets, enhance predictive accuracy, and provide flexibility and interpretability makes it a valuable tool in actuarial science.

1.5. The Novelty for Application of the Extreme Machine Learning Algorithm in This Study

The application of the Extreme Machine Learning (XML) algorithm in this study represents a novel approach to revolutionize the underwriting process in agricultural insurance. XML techniques, including deep neural networks, ensemble methods, and reinforcement learning, offer unique advantages in handling the complexity and heterogeneity of agricultural data, enabling more accurate risk assessment and dynamic underwriting decisions [17]. XML algorithms excel in capturing intricate patterns and relationships within vast datasets, such as historical weather patterns, soil quality indices, crop yield data, market trends, and socio-economic factors, which are essential for comprehensive risk profiling in agricultural insurance [18]. By leveraging these algorithms, the proposed model can overcome the limitations of traditional underwriting approaches, which often struggle to adapt to evolving environmental conditions and market dynamics [19].

Furthermore, the integration of XML techniques with actuarial intelligence enhances the model's robustness, transparency, and scalability, ensuring alignment with industry standards and regulatory requirements. This synergy between advanced machine learning methods and actuarial principles enables the development of a sophisticated underwriting model that can effectively address the unique challenges and opportunities in agricultural insurance.

1.6. The Novelty of the Study

Integration of multiple methodologies: The code combines various methodologies including statistical simulation, machine learning, actuarial modeling, and risk analysis within a unified framework. This interdisciplinary approach contributes novelty by offering a comprehensive solution to agricultural insurance challenges. The segmentation of policyholders into underwriting bands based on claim amounts introduces a novel way to assess risk exposure and pricing strategies. This categorization facilitates a more targeted and customized approach towards risk management and pricing. The incorporation of robustness, stress, and scenario testing methodologies enhances the robustness of the study. These tests provide insights into the stability and resilience of the developed models and pricing strategies under different conditions, contributing novel insights into risk management practices.

1.7. Contribution to the Body of Knowledge

Advancement in agricultural insurance modeling: The study contributes to the advancement of agricultural insurance modeling techniques by integrating modern data simulation, machine learning, and actuarial methodologies. This contributes to the broader body of knowledge within the actuarial science and insurance domains. By automating actuarial risk pricing processes using machine learning models, the study contributes towards enhancing efficiency and accuracy in insurance operations. This can lead to improved risk management practices and better financial outcomes for insurers and policyholders. The insights derived from exploratory data analysis, modeling, and testing provide valuable information for decision-making in agricultural insurance underwriting, pricing, and risk management. This contributes to informed decision-making and improved sustainability of agricultural insurance products.

2. Review of Methods

Previous studies have explored various aspects of agricultural insurance, machine learning applications, and actuarial science. [11] conducted a comprehensive review of machine learning applications in agricultural insurance, highlighting the potential for innovation and advancement in risk management strategies. Similarly, [12] discussed advancements in agricultural risk modeling, emphasizing the role of machine learning techniques in improving risk assessment accuracy. Additionally, [19] explored the integration of machine learning and actuarial science

in agricultural insurance, emphasizing the importance of aligning predictive modeling with industry standards and regulatory requirements.

The proposed study employs a multi-faceted approach to develop the Extreme Machine Learning Actuarial Intelligent Agricultural Insurance Based Automated Underwriting Model. The methodology involves data collection from diverse sources, including historical weather data, soil quality indicators, crop yield records, market trends, and socio-economic factors. Machine learning techniques, such as deep neural networks, ensemble methods, and reinforcement learning, are utilized to analyze and process the data for risk assessment purposes. Actuarial principles are integrated into the model to ensure alignment with industry standards and regulatory compliance. Model validation techniques, including cross-validation and sensitivity analysis, are employed to assess the robustness and reliability of the underwriting model. The study adopts a dynamic underwriting approach, enabling real-time adjustments to changing environmental conditions, market dynamics, and policyholder characteristics.

3. Methodology

This provides an overview of the problem statement, emphasizing the need for advanced modeling techniques in agricultural insurance underwriting. Introduce the concept of Extreme Machine Learning (ELM) as the primary modeling approach.

3.1. Data Generation

Describe the process of simulating agricultural insurance data incorporating various factors such as climate change scenarios, crop types, farm sizes, continuous and binary variables, and numerical variables. Generate a large dataset representative of agricultural insurance scenarios to facilitate model development and testing.

3.2. Exploratory Data Analysis (EDA)

Conduct EDA to understand the characteristics and distributions of the simulated data. Perform summary statistics, frequency tables, and visualization (histograms and bar plots) to gain insights into the data's structure and relationships.

3.3. Hypothesis Testing

Formulate hypotheses related to agricultural risk factors. Utilize statistical tests such as ANOVA to investigate significant differences between variables, e.g., crop yield across different farm sizes.

3.4. Data Partitioning, Model Building, and Feature Engineering

Split the dataset into training and testing sets using the 80:20 rule. Implement Extreme Learning Machines (ELM) for model building, considering variables such as crop yield, loss ratio, pest infestation, drought, temperature, rainfall, soil moisture,

etc. Conduct feature engineering if necessary, such as scaling or encoding categorical variables.

3.5. Model Validation

Evaluate model performance using metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Validate the model's predictions against actual data, adjusting for factors like inflation rates.

3.6. Actuarial Risk Pricing Model

Develop separate models for frequency risk, severity risk, and inflation adjustment risk. Utilize ELM to estimate premiums based on various agricultural risk factors. Integrate the models to automate actuarial risk pricing for agricultural insurance policies.

3.7. Actuarial Loss Reserve Risk Premium Balances

Predict case reserves using the ELM model and adjust for inflation. Calculate loss reserving risk premium balances by combining predicted case reserves with automated premiums.

3.8. Actuarial Underwriting Bands

Define underwriting bands based on automated premiums, loss reserves, and risk premium balances. Categorize policyholders into different underwriting bands to assess risk levels.

3.9. Segment Policyholders Based on Their Underwriting Band and Average Claim Amount

Analyze the distribution of policyholders across underwriting bands and their corresponding claim amounts. Ensure compliance with International Financial Reporting Standard 17 (IFRS17) regulations.

3.10. Robust Tests, Stress Tests, and Scenario Tests

Perform robustness tests by varying inflation rates and evaluating the impact on loss reserving risk premium balances; conduct stress tests by increasing claim amounts to simulate adverse scenarios and assess the resilience of the model; execute scenario tests by introducing specific changes in the data and observing the model's response.

4. Data

Here is the simulation process and general description of the simulated Agricultural Insurance data used for developing an Automated Actuarial Underwriting Model with a sample size of 100,000 policyholders.

Policyholder ID: This variable represents a unique identifier for each policyholder. It doesn't directly contribute to the model but is useful for tracking and

identifying individual policyholders. *Climate Change:* Simulates different scenarios of climate change, namely "No Change", "Moderate Change", and "Severe Change". This variable is crucial as it affects various aspects of agricultural production and risk, such as temperature, rainfall, and soil moisture. *Crop Type:* Represents the type of crop being cultivated. Different crops have varying sensitivities to environmental factors and risks, making it essential to consider when assessing insurance risk. *Farm Size:* Indicates the size of the farm, which can influence the scale of production and susceptibility to certain risks. Larger farms may have different risk profiles compared to smaller ones. *Crop Yield:* Simulates the yield of the crop in kilograms. Crop yield is a fundamental factor affecting agricultural insurance, as it directly impacts the potential revenue and losses for the policyholder. *Loss Ratio:* Represents the ratio of losses to premiums collected. It provides insight into the profitability and risk exposure of the insurance policy. *Pest Infestation and Drought:* Binary variables indicating whether the farm is affected by pest infestation or drought. These are common risks in agriculture and can significantly impact crop yields and profitability. *Temperature, Rainfall, and Soil Moisture:* These variables simulate climatic conditions, which are crucial determinants of crop health and productivity. Temperature, rainfall, and soil moisture levels directly influence crop growth and vulnerability to pests and diseases. *Number of Claims and Claim Amount:* Simulate the number and amount of insurance claims filed by policyholders. These variables reflect the actual losses experienced by farmers due to various risks covered by the insurance policy. *Number Premium of Payments and Agricultural Premium*: Represent the number of premium payments made by policyholders and the corresponding agricultural insurance premiums. These variables are essential for calculating the revenue and financial sustainability of the insurance company. *Inflation Rate:* Simulates inflation rates, which can affect the value of premiums, claims, and reserves over time. Accounting for inflation is crucial for ensuring the financial stability and adequacy of insurance reserves. *Case Reserves:* Represents the reserves set aside by the insurance company to cover potential future claims. It reflects the financial strength and risk management practices of the insurer.

In short, the simulated data cover a wide range of factors relevant to agricultural insurance, including climate conditions, crop characteristics, risk events, financial metrics, and policyholder behavior. By analyzing these variables, insurers can develop predictive models and underwriting strategies to accurately assess risk, price premiums, and manage their portfolios effectively.

5. Results

This section presents the outcomes of this study.

5.1. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a crucial step in the data analysis process, involving the examination and visualization of data to understand its underlying

patterns, characteristics, and relationships. EDA techniques help analysts gain insights into the structure and distribution of the data, identify outliers and missing values, and inform subsequent modeling decisions [20]. In agricultural insurance, EDA plays a vital role in understanding the various factors influencing risk and loss outcomes. By exploring the dataset through EDA, analysts can uncover meaningful associations between climate conditions, crop types, farm characteristics, and insurance outcomes. This exploration aids in the formulation of hypotheses and informs the development of predictive models for loss reserves, risk pricing models and the underwriting models.

5.1.1. Exploratory Data Analysis for Continuous Variables in the Simulated Data

The Exploratory Data Analysis (EDA) for continuous variables in the simulated agricultural insurance data involves examining the distribution and summary statistics of numerical attributes to gain insights into their characteristics and potential relationships.

Figure 3(a) shows the distribution of crop yield in kilograms across the dataset. It gives an idea of the range and distribution of crop yields among policyholders. Higher values indicate higher crop yields. **Figure 3(b)** displays the distribution of loss ratios. It represents the proportion of losses to premiums collected. Higher values indicate higher proportions of losses relative to premiums, suggesting poorer underwriting performance. **Figure 3(c)** depicts the distribution of temperatures and it shows the range and frequency of temperatures experienced in the areas where policyholders operate farms. This variable is crucial for assessing climate-related risks. **Figure 3(d)** illustrates the distribution of rainfall levels and it indicates the range and frequency of rainfall amounts, which is essential for understanding water availability and drought risk. **Figure 3(e)** shows the distribution of soil moisture levels. Moreover, it reflects the moisture content of the soil, which is vital for crop growth and susceptibility to drought or water logging. **Figure 3(f)** displays the frequency distribution of the number of claims filed by policyholders. It provides insight into the frequency of insurance claims, which is essential for assessing risk exposure and loss potential. **Figure 3(g)** illustrates the distribution of claim amounts and it shows the range and frequency of claim amounts, indicating the severity of losses experienced by policyholders. **Figure 3(h)** depicts the distribution of the number of premium payments made by policyholders and it provides insight into the payment behavior of policyholders and their commitment to maintaining insurance coverage. **Figure 3(i)** shows the distribution of agricultural insurance premiums and it indicates the range and frequency of premium amounts paid by policyholders, reflecting the cost of insurance coverage. **Figure 3(j)** illustrates the distribution of inflation rates and it shows the range and frequency of inflation rates, which can impact the value of premiums, claims, and reserves over time. **Figure 3(k)** displays the distribution of case reserves. It represents the amount of funds set aside by the insurance company to cover potential future

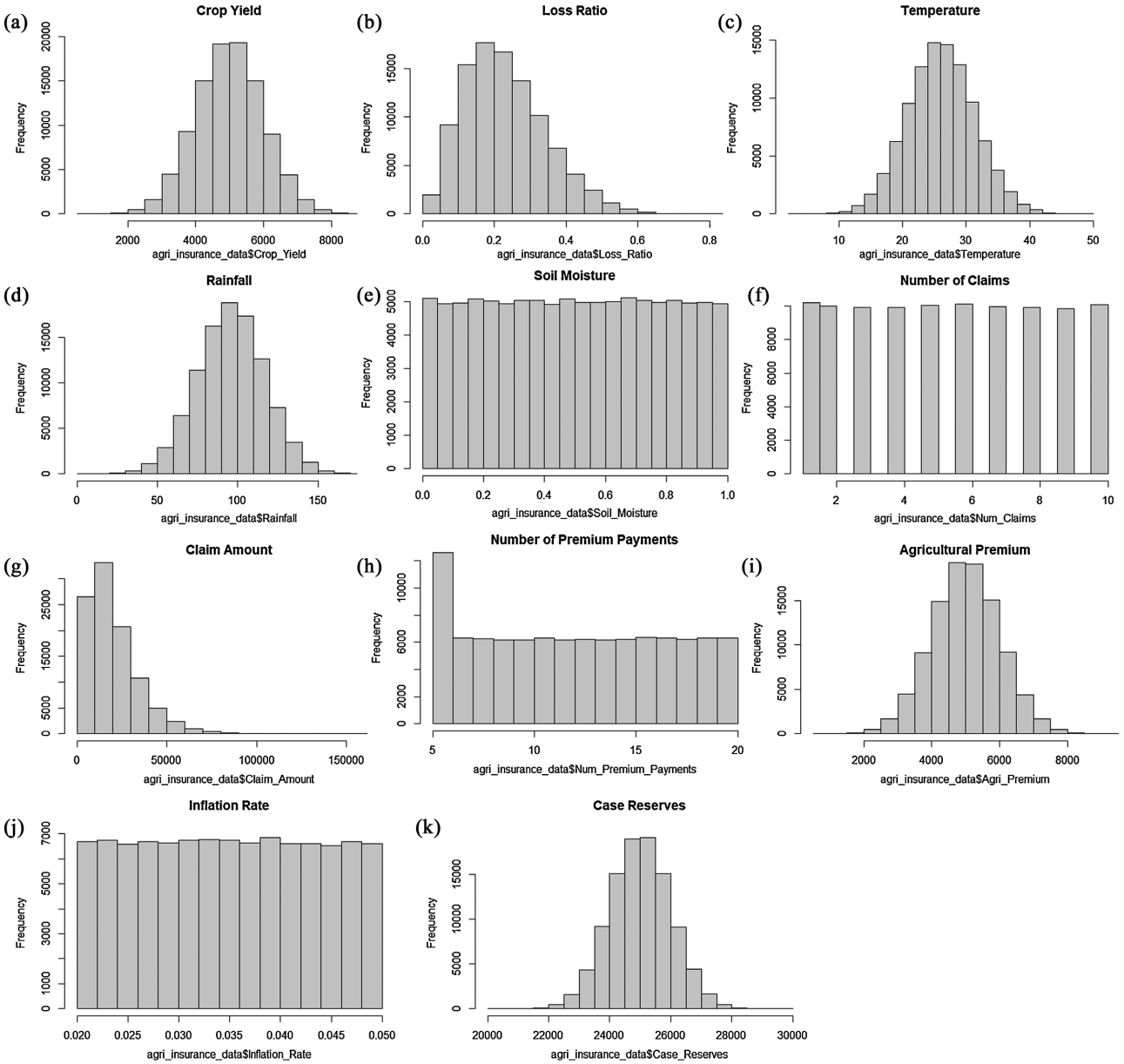

Figure 3. (a) Crop Yield; (b) Loss Ratio; (c) Temperature; (d) Rainfall; (e) Soil moisture; (f) Number of claims; (g) Claim Amount; (h) Number of premium payments; (i) Agricultural Premium; (j) Inflation Rate; (k) Case Reserves.

claims. Higher values indicate higher reserve levels, reflecting the insurer's risk management practices.

5.1.2. Exploratory Data Analysis for Categorical Variables

In general, the EDA for categorical variables helps identify the composition and prevalence of different categorical attributes in the dataset, providing essential insights into the diversity of agricultural operations and risk exposures among policyholders. These insights are valuable for informing subsequent modeling and analysis tasks in agricultural insurance underwriting and risk management.

Figure 4(a) indicates the frequency distribution of different climate change

scenarios among policyholders. It shows the proportion of policyholders experiencing "No Change," "Moderate Change," and "Severe Change" in climate conditions. The majority of policyholders experience "No Change," while fewer are affected by "Moderate Change" or "Severe Change" in climate conditions. Figure 4(b) illustrates the distribution of different crop types grown by policyholders. It shows the prevalence of each crop type (e.g., Wheat, Corn, Soybean, Rice) among policyholders. Some crops may be more commonly grown than others, reflecting the diversity of agricultural practices. Figure 4(c) displays the frequency distribution of different farm sizes. It indicates the prevalence of small, medium, and large farms among policyholders. The distribution of farm sizes reflects the variability in the scale of agricultural operations within the dataset. Figure 4(d) illustrates the occurrence of pest infestation among policyholders. It shows the proportion of policyholders affected by pest infestation. Some policyholders experience pest infestations, while others do not, indicating varying levels of pest management practices or environmental conditions. Figure 4(e) indicates the occurrence of drought among policyholders. It shows the proportion of policyholders affected by drought events. Similar to pest infestation, some policyholders experience drought, while others do not, reflecting the variability in drought risk exposure across agricultural operations.

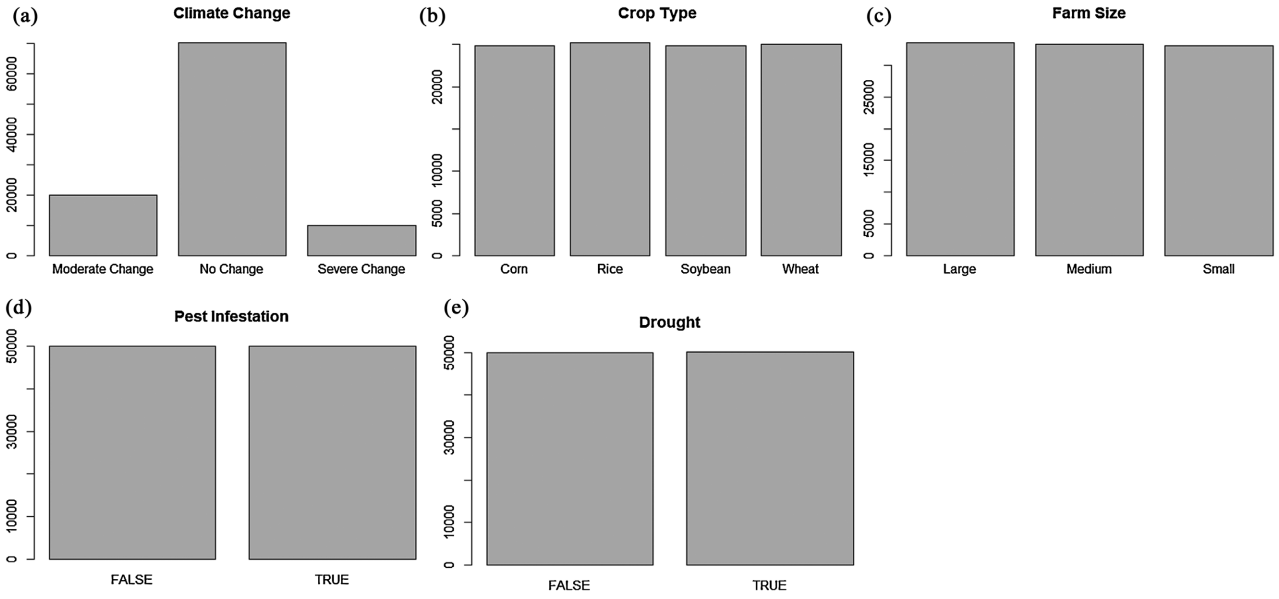

Figure 4. (a) Climate change scenarios; (b) Distribution of crop types; (c) Distribution of farm sizes; (d) Occurrence of pest infestations; (e) Occurrence of drought.

5.1.3. ANOVA Test between Farm Size and Crop Type

ANOVA, or Analysis of Variance, is a statistical method used to analyze the differences among group means in a sample. It is commonly employed when comparing three or more group means simultaneously. Moreover, the ANOVA partitions the total variance observed in the data into different sources of variation, namely the variation between groups and the variation within groups. By comparing

the ratio of between-group variation to within-group variation, ANOVA assesses whether the differences in group means are statistically significant or if they could have occurred by random chance alone [21]. ANOVA is particularly useful in experimental settings where researchers want to determine if the means of several groups differ significantly from each other. It is widely applied across various disciplines, including psychology, biology, economics, and social sciences, to analyze experimental data with multiple treatment conditions or categorical factors [22].

In this case, Null Hypothesis (H_0): There is no significant difference in crop yield between different farm sizes. Alternative Hypothesis (H_1): There is a significant difference in crop yield between different farm sizes.

Table 1. ANOVA table of results.

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
Farm Size	2	172168.02	86084.01	0.09	0.9171
Residuals	99,997	99497616806.37	995006.02		

The *p*-value (0.9171) for Farm Size, from **Table 1** is much greater than the typical significance level of 0.05. Therefore, we fail to reject the null hypothesis. This suggests that there is no significant effect of Farm Size on Crop Yield, as the *p*-value is not less than 0.05. Thus, based on the ANOVA results, we do not have sufficient evidence to conclude that farm size has a significant effect on crop yield.

. Estimation of Actuarial Case Loss Reserves using predictive model. To build an Extreme Learning Machine (ELM) based Actuarial Inflation-adjusted Case Reserve estimation model, we begin by loading the required libraries, including *elmNNRcpp*, proceed to partition the data into training and testing sets using an 80:20 split; then preprocess the data if necessary. Define the model architecture and train the ELM model and adjust the Case Reserves using inflation rates.

In the context of agricultural insurance, predictive modeling for loss reserves would typically involve building regression models or machine learning algorithms that leverage variables such as climate conditions, crop types, farm characteristics, and historical claims data to predict future losses. These models may use techniques like linear regression, generalized linear models (GLMs), decision trees, random forests, or neural networks to capture complex relationships between predictors and loss outcomes [22].

Table 2 is a summary of the results obtained from the ELM model and each row of the table represents a different attribute or component of the model. *inpweight* indicates the input weights of the model. Input weights represent the coefficients assigned to each predictor variable in the model. However, since ELM models use a random initialization of input weights, these values may not provide direct interpretation. *biashid* indicates the bias term for the hidden layer of the model. The bias term is added to the weighted sum of inputs before passing through the activation function of the hidden layer. *outweight* indicates the output

weights of the model. Output weights represent the coefficients assigned to the hidden layer's output neurons in predicting the target variable. *actfun* indicates the activation function used in the hidden layer of the model. Activation functions introduce non-linearity into the model and enable it to learn complex patterns in the data. *nhid* indicates the number of hidden neurons in the model. The number of hidden neurons determines the model's capacity to learn and represent complex relationships in the data. *predictions* indicates the predicted values obtained from the model for the training data. *fitted values* This indicates the fitted values obtained from the model for the training data. Fitted values represent the model's predictions based on the input data. *residuals* indicates the residuals or errors obtained from the model's predictions for the training data. Residuals represent the differences between the observed and predicted values. *formula* indicates the formula used to specify the model. It includes the response variable and predictor variables used in the model. *call* indicates the call or function used to create the model. *is regression* indicates whether the model is a regression model (logical value). *is logical* indicates whether the model is a logical model (logical value).

Table 2. Summary results for ELM based Actuarial loss reserve estimation.

	Length	Class	Mode
inpweight	220	-none-	numeric
biashid	0	-none-	numeric
outweight	20	-none-	numeric
actfun	1	-none-	character
nhid	1	-none-	numeric
predictions	80,000	-none-	numeric
fitted_values	80,000	-none-	numeric
residuals	80,000	-none-	numeric
formula	3	terms	call
call	5	-none-	call
is_regression	1	-none-	logical
is_logical	1	-none-	logical

Shortly, this summary provides information about the components and attributes of the ELM model, including input weights, biases, activation function, number of hidden neurons, predictions, residuals, and model formula. However, since ELM models are primarily used for prediction rather than interpretation, the specific values of input weights and biases may not be directly interpretable in the same way as traditional regression coefficients.

Figure 5 visualizes the relationship between the actual loss reserves (Case Reserves) and the predicted loss reserves (Inflation Adjusted Reserves) generated by your Extreme Machine Learning (ELM) model. The X-axis (Actual Case Loss

Reserves) represents the actual values of loss reserves, which are the amounts set aside by the insurance company to cover potential claims from policyholders. Moreover, each point on the x-axis corresponds to the actual loss reserves for a specific observation in the simulated test data set. The Y-axis (Predicted Loss Reserves) represents the predicted values of loss reserves generated by the ELM model. These are the estimates of loss reserves made by your model based on the input features such as crop yield, loss ratio, weather conditions, etc. Once again, each point on the y-axis corresponds to the predicted loss reserves for a specific observation in the simulated test data set.

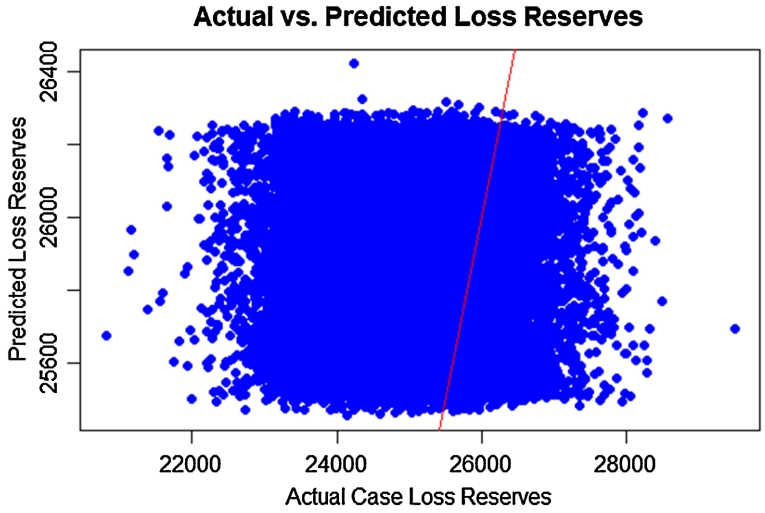

Figure 5. Actual case reserves vs predicted loss reserves.

There is evidence of the points above the red line indicate instances where the model overestimated the loss reserves compared to the actual values and on the same note there are points below the red line which indicate instances where the model underestimated the loss reserves compared to the actual values. However, the red line passes through the majority of blue points. This is due to the *rnorm*() which has been used to simulate the case reserves. In general, there is a model prediction accuracy as a result of the closeness of the points to the red line.

5.3. Traditional Chain Ladder Model

The Chain Ladder method operates on the assumption that historical claims development patterns will continue into the future. It involves creating a chain of estimated development factors for each period between the valuation date and the ultimate development of claims. These factors are then applied to known or estimated claims data to project future payments. The method often employs techniques like weighted averages or simple ratio methods to estimate these factors [18].

5.3.1. Structure of General Chain Ladder

The Chain Ladder method typically follows these steps: Organize historical claims data by accident period and development period. Calculate development factors

for each development period based on historical data. Common methods include age-to-age factors, average factors, or other weighted averages. Projection of Outstanding Claims by applying the development factors to the known incurred losses to estimate the ultimate losses for each accident period. Reserve Calculation by summing up the estimated ultimate losses for each accident period to obtain the total outstanding claims reserves.

Table 3. General chain ladder triangle.

		Development Periods			
Accident Years		1	2	⋯	n
		I_{11}	I_{12}	⋯	I_{1n}
		I_{21}	I_{22}	⋯	I_{2n}
⋮	⋮	⋮	⋮	⋱	⋮
m		I_{m1}	I_{m2}	⋯	I_{mn}

With illustration from Table 3, let I_{ij} represent the incurred claims for accident period i at development period j.

$$DF_{ij} = \frac{I_{i,j+1}}{I_{ij}} \tag{20}$$

where DF_{ij} are the estimated development factors.

Projection of Outstanding Claims is carried out by Equation (21) below

$$\text{Ultimate Loss}_i = I_{i1} \times \prod_{j=1}^{n} DF_{ij} \tag{21}$$

The Reserve Calculation is carried out by (22)

$$\text{Total Reserves} = \sum_{i=1}^{m} \text{Ultimate Loss}_i \tag{22}$$

5.3.2. Simulated Run-Off Triangle

The Traditional chain ladder based on the simulated agricultural insurance data has been presented below.

Figure 6 is a graphical representation of the development of claims over time, usually separated by accident or underwriting years and development periods. The x-axis represents the development period, often corresponding to different durations after the accident or underwriting year and the y-axis represents the cumulative amount of claims incurred over time. Each line in the plot represents the development of claims for a specific accident or underwriting year. The lines start from the origin (0, 0) and extend as the claims develop over subsequent periods. The slope of each line indicates the pattern of claims development. Steeper slopes indicate faster development of claims, while flatter slopes indicate slower development. Figure 7 is similar to the regular claims development plot but may offer additional features depending on the specific lattice settings. Both types of

plots are commonly used in actuarial analysis to visualize and analyze the development of insurance claims over time, helping actuaries understand trends, identify outliers, and assess the adequacy of reserves.

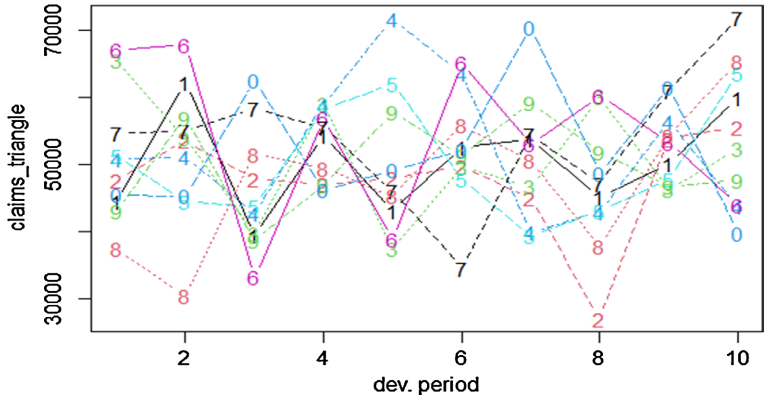

Figure 6. Simulated claims development profile.

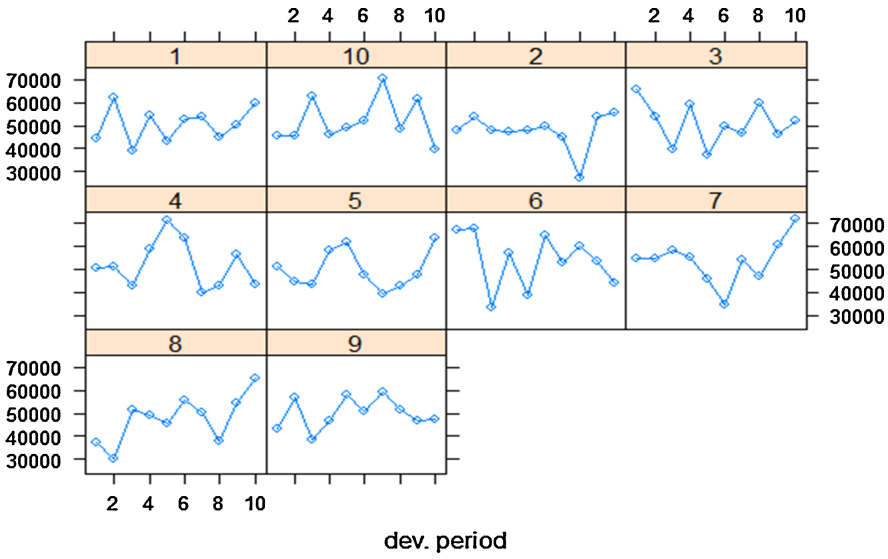

Figure 7. Regular claims development plot.

5.4. Comparison between Simulated Traditional Chain Ladder Model and the ELM Based Loss Reserving Period

The two methods, the Traditional Chain Ladder Model and the ELM (Extreme Machine Learning) based Loss Reserving method, were compared in terms of their effectiveness in estimating reserves for agricultural insurance.

The Traditional Chain Ladder Model estimates reserves using a method based on historical patterns of claims development. It typically calculates reserves by extrapolating past claims data and projecting future claims based on past experience. The ELM based Loss Reserving method utilizes a machine learning algorithm, specifically Extreme Learning Machines (ELM), to predict future claims and estimate reserves. This method learns complex patterns and relationships from the data to make predictions.

Both methods were applied to the same dataset of simulated agricultural insurance data. This dataset contained various variables such as climate conditions, crop types, farm sizes, historical claims, premiums, and other relevant factors.

For the Chain Ladder Model, the historical claims data were structured into a claims development triangle. The model then used this triangle to extrapolate future claims and estimate reserves. For the ELM based Loss Reserving method, the machine learning model was trained using the features available in the dataset to predict future claims and estimate reserves. The Chain Ladder Model and the ELM model independently calculated reserve estimates based on the provided data and their respective methodologies. Once the reserve estimates were obtained from both methods, they were compared directly using a barplot visualization. The barplot displayed the estimated reserves for each method side by side, allowing for easy visual comparison. The height of each bar represented the magnitude of reserves estimated by each method, providing a clear indication of which method yielded higher or lower reserve estimates. Additionally, text labels on top of each bar provided precise numerical values for further comparison.

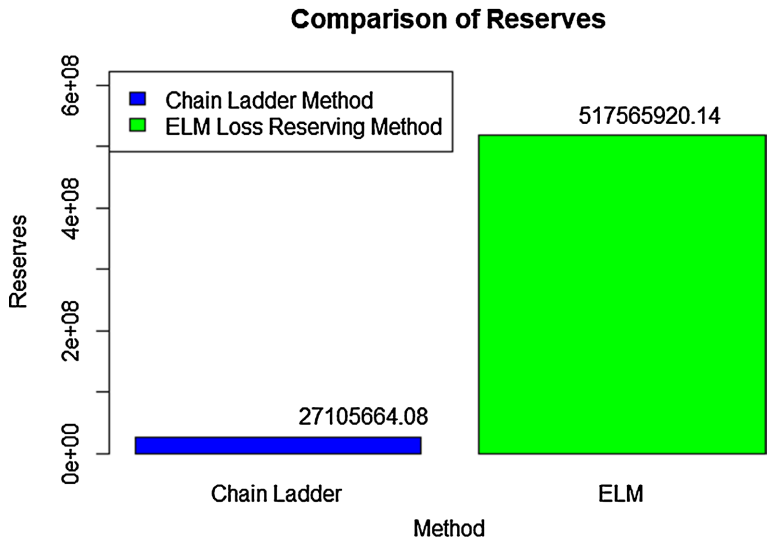

Figure 8. Simulated chain ladder vs ELM loss reserving models.

Figure 8 compares the reserves estimated by two different methods: the Traditional Chain Ladder Model and the Extreme Machine Learning (ELM) based Loss Reserving method. Two bars are shown, one for each method, indicating the amount of reserves estimated by each method. The blue bar represents the reserves estimated using the Traditional Chain Ladder Model and the green bar represents the reserves estimated using the ELM based Loss Reserving method. In addition to that, the comparison aims to assess the effectiveness and accuracy of the two methods in estimating reserves for agricultural insurance. The height of each bar indicates the magnitude of reserves estimated by each method and by comparing the heights of the bars, the green bar (ELM method) is higher than the blue bar (Chain Ladder method), it suggests that the ELM method estimates

higher reserves compared to the traditional Chain Ladder approach. Hence this makes the ELM Loss Reserving method better than the Traditional chain ladder method.

5.5. Determination of the Actuarial Risk Premiums

The automated risk premiums in this paper are determined through a process involving three separate models:

- *Frequency Risk Pricing Model:* This model predicts the frequency of premium payments based on various factors such as crop yield, crop type, loss ratio, pest infestation, drought, temperature, rainfall, soil moisture, number of claims, and claim amount. It estimates how often a policyholder is expected to make premium payments.
- *Severity Risk Pricing Model:* This model predicts the severity of potential losses (*i.e.*, the size of potential claims) based on similar factors as the frequency model. It estimates the financial impact of each claim.
- *Inflation Adjustment Risk Pricing Model:* This model predicts the inflation rate based on the same set of factors. It estimates the rate at which prices are expected to rise over time.

Once these three models are trained using historical data, they can be used to make predictions for new policyholders. The predicted frequency, severity, and inflation rate are then combined to calculate the automated actuarial risk premiums for each policyholder.

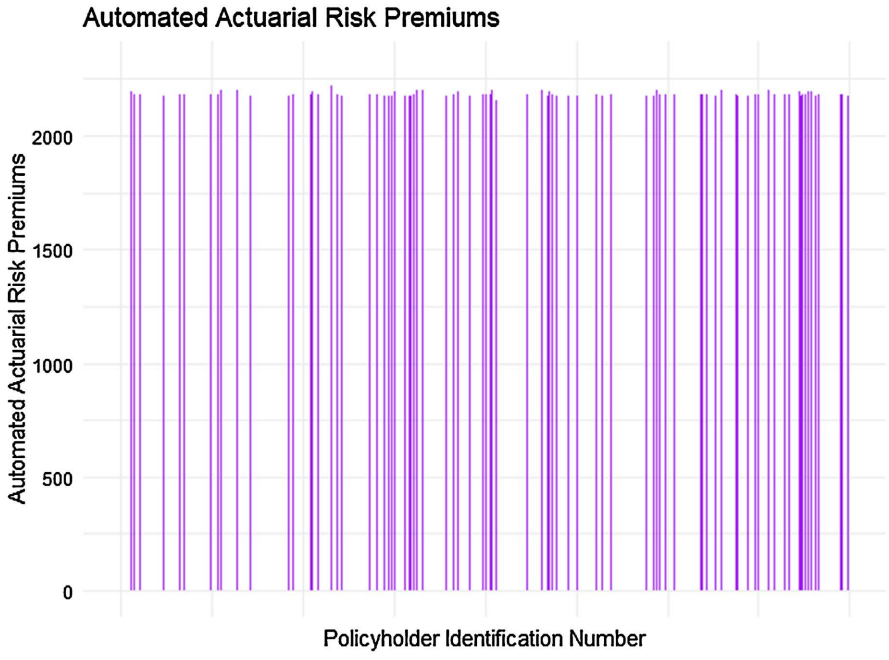

Figure 9. Automated actuarial risk premiums.

The X-axis is the Policyholder Identification Number from Figure 9 which is used to uniquely identify each policyholder and the Y-axis is the Automated

Actuarial Risk Premiums, which represents the calculated risk premiums for each policyholder based on the automated modeling process. The height of each bar represents the magnitude of the automated actuarial risk premium assigned to each policyholder. Taller bars indicate higher risk premiums, while shorter bars indicate lower premiums. In this case, all the bars are taller, however this draws the need to create underwriting bands to alleviate the riskiness among the policyholders and this has been shown Subsection 5.7.

5.6. Actuarial Loss Reserving Risk Pricing Balances

The actuarial loss reserve risk premium balances have been developed by combining two key components: First, the inflation-adjusted loss reserves are calculated based on the predicted case reserves obtained from the ELM (Extreme Machine Learning) model and adjusting them for inflation. This has been illustrated on Subsection 5.2. These predicted case reserves are then adjusted for inflation by multiplying them with the inflation rates present in the test data. Secondly, Automated actuarial risk premiums are calculated as illustrated by Subsection 5.5 above. Finally, the inflation-adjusted loss reserves and automated actuarial risk premiums are summed together to form the actuarial loss reserve risk premium balances. This combined metric represents the total expected liabilities and risk exposure for the insurance company, incorporating both predicted future claim reserves and risk premiums.

Figure 10. Actuarial loss reserving risk pricing balances.

The height of each bar on **Figure 10** corresponds to the Actuarial Loss Reserve Risk Premium Balance for the respective policyholder. Higher bars indicate higher risk exposures or larger financial obligations. By observing the Figure above, you can

identify policyholders with higher Actuarial Loss Reserve Risk Premium Balances, indicating potentially higher risk exposures or financial obligations. It helps insurers in understanding the distribution of risk across their policyholder base and in making informed decisions related to risk management, pricing, and financial planning.

5.7. Creating Actuarial Underwriting Bands

This section describes how R code employed to both calculate and visualize four underwriting bands based on different criteria, including automated actuarial risk premiums, inflation-adjusted loss reserves, and a randomly generated reinsurance amount. Here is an explanation of how the underwriting bands were formed respectively.

1) *Underwriting Band 1:* This band is based on the range of Automated Actuarial Risk Premiums (AARP). The minimum value of the automated actuarial risk premiums is calculated using min(AARP), and the maximum value is calculated using max(AARP). This band represents the range of premiums generated automatically through the actuarial risk pricing model.

2) *Underwriting Band 2:* This band is based on the range of Inflation-Adjusted Loss Reserves (IALR). The minimum value of the inflation-adjusted loss reserves is calculated using min(IALR), and the maximum value is calculated using max(IALR). This band represents the range of reserves adjusted for inflation.

3) *Underwriting Band 3:* This band is based on the range of Actuarial Loss Reserving Risk Premium Balances (ALRRPB). The minimum value of the combined balances is calculated using min(ALRRPB), and the maximum value is calculated using max(ALRRPB). This band represents the range of combined balances after adjusting for inflation and including Automated Actuarial Risk Premiums.

4) *Underwriting Band 4:* This band is generated randomly using the *runif*() function. It represents a range of potential reinsurance amounts, ranging from the maximum value of the combined balances (max(ALRRPB)) to a maximum value of $1,000,000. This band introduces variability in potential reinsurance amounts.

Figure 11 visualizes these underwriting bands, with each band represented by a different color. The x-axis represents the categories of underwriting bands, while the y-axis represents the maximum value of each band. Error bars are used to represent the range of values within each band.

Policyholder Categorization According to the Created Actuarial Underwriting Bands

The policyholders have been categorized into four underwriting bands based on their claim amounts. Each band has a specific range of claim amounts associated with it. These bands are defined based on thresholds determined from the data as explained on the earlier subsection 5.7. An R code has been developed which defines a function named *categorize_underwriting_band* (*claim_amount*) which takes the claim amount as input and categorizes the policyholder into one of the

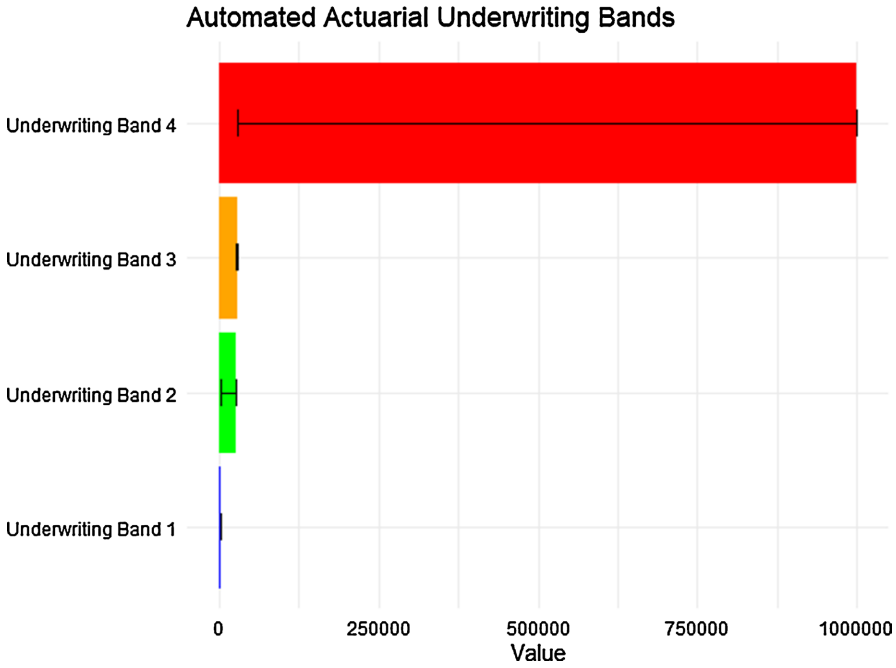

Figure 11. Actuarial underwriting bands.

underwriting bands based on their claim amount. From there the *sapply*() function is used to apply the *categorize_underwriting_band*() function to each claim amount in the data set. This results in each policyholder being assigned to one of the underwriting bands based on their claim amount. Moreover, the code creates a new variable named "Underwriting_Band" in the dataset, which stores the underwriting band assigned to each policyholder based on their claim amount. Finally, the code then visualizes the distribution of policyholders across the underwriting bands and also shows the average claim amount within each band. This visualization 26 helps in understanding how policyholders are distributed across different risk categories based on their claim amounts.

Figure 12 displays the number of policyholders in each underwriting band, represented by the height of the bars. Additionally, it shows the average claim amount within each band, represented by the red points and connecting lines. Policyholders with higher claim amounts are typically assigned to higher underwriting bands, indicating higher risk levels. Figure 12 provides insights into the distribution of policyholders based on their risk profiles, allowing insurers to assess their exposure to different levels of risk and make informed underwriting decisions. Adherence to IFRS17 Regulations, Figure 12 visualizes the distribution of policyholders across different underwriting bands based on their claim amounts, along with the average claim amount within each band. Here is how this adherence to IFRS17 regulations, respectively.

- *Risk Classification:* IFRS17 requires insurers to classify insurance contracts into groups with similar risk profiles. In this plot, policyholders are categorized into underwriting bands based on their claim amounts, which serves as a proxy for their risk exposure. This classification aligns with the principles of

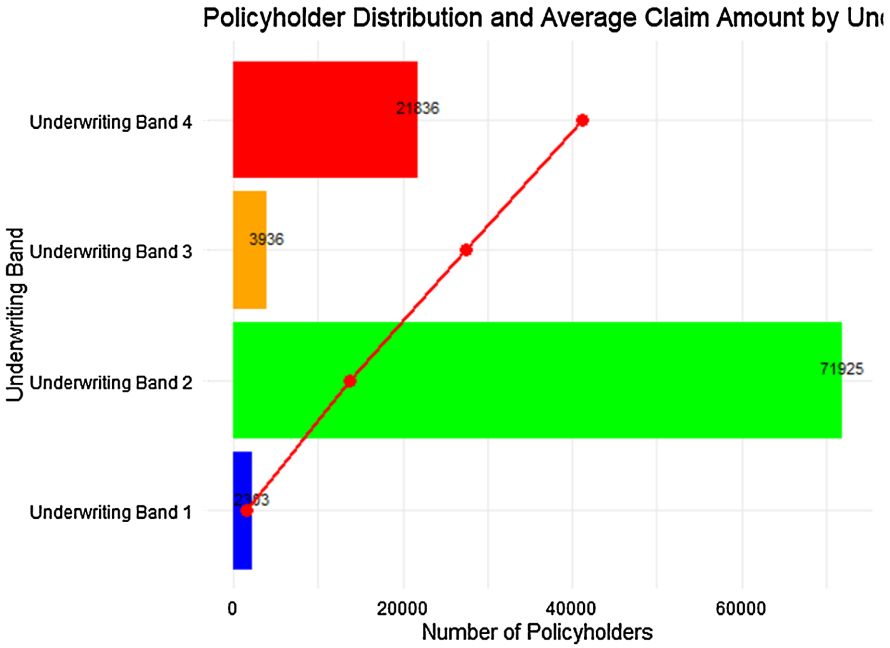

Figure 12. Policyholder categorization based on the created actuarial underwriting bands.

risk segmentation mandated by IFRS17.

- *Transparency and Disclosure:* IFRS17 emphasizes transparency and disclosure in financial reporting. By visualizing the distribution of policyholders across underwriting bands, insurers can provide stakeholders with a clear understanding of the risk composition of their insurance portfolio. This transparency helps in accurately assessing the financial position and performance of the insurer.
- *Actuarial Assumptions and Estimates:* Insurers under IFRS17 are required to make various actuarial assumptions and estimates, including claims reserves. The visualization of average claim amounts within each underwriting band provides insight into the actuarial estimation process. It demonstrates how claim amounts vary across different risk categories, enabling insurers to make informed decisions regarding reserve calculations and financial disclosures.
- *Fair Value Measurement:* IFRS17 introduces the concept of fair value measurement for insurance contracts. By categorizing policyholders into underwriting bands based on claim amounts, insurers can better assess the fair value of their insurance liabilities. This classification helps in determining appropriate reserve levels and pricing strategies, ensuring that insurers accurately reflect the value of their insurance contracts on financial statements.

In closing, Figure 12 adheres to IFRS17 regulations by facilitating risk classification, promoting transparency and disclosure, supporting actuarial assumptions and estimates, and enabling fair value measurement of insurance contracts. It provides valuable insights into the risk profile of the insurance portfolio, aiding insurers in meeting the requirements of IFRS17 and enhancing financial reporting practices.

5.8. Model Evaluation: Robust Tests, Stress Tests and Scenario Tests

In short, model evaluation through robust tests, stress tests, and scenario tests plays a critical role in determining actuarial loss reserve risk balances. These tests provide insurers with valuable insights into the reliability, resilience, and potential vulnerabilities of their reserve models, enabling them to make informed decisions and manage risks effectively [23].

5.8.1. Robust Tests

Robustness tests ensure that the model is stable and performs consistently under different conditions. We can conduct robustness tests by checking the stability of the model's predictions under variations in input parameters [23]. Let us consider the stability of the loss reserving risk pricing balances across different levels of inflation rates.

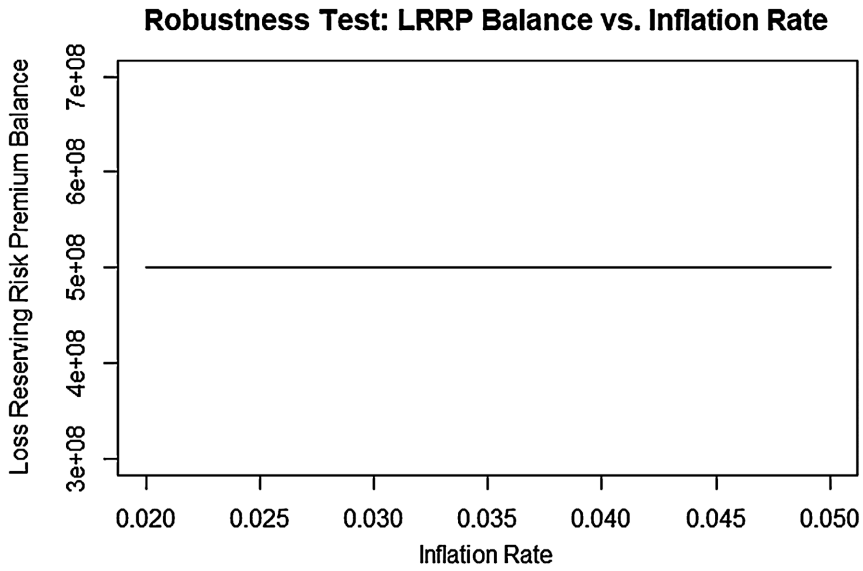

Figure 13. Robust test plot.

Figure 13. shows a relatively consistent trend or pattern as inflation rates vary, it suggests that the model's predictions are robust across different inflation scenarios. In other words, the model's performance remains stable even when faced with changes in inflation rates. Thus, the plotted line remains relatively flat, it indicates that the model's predictions for Loss Reserving Risk Premium Balances are consistent across different inflation rates. This consistency is a sign of robustness.

5.8.2. Stress Tests

Stress tests evaluate the resilience of the model under extreme scenarios or adverse conditions. We can simulate extreme scenarios and observe the model's response [24]. Let us stress test the model by introducing a significant increase in claim amounts.

Figure 14. Baseline vs. stress test plot.

Figure 14 compares the Loss Reserving Risk Premium Balance between two scenarios: the baseline (original) and the stress test (where claim amounts are increased by a factor of 1.5). The baseline represents the Loss Reserving Risk Premium Balance under normal conditions, where claim amounts are not altered and the Stress Test reflects the Loss Reserving Risk Premium Balance under the stress scenario, where claim amounts are increased by a factor of 1.5. In this case, the stress test balances the baseline, it indicates that the model is responsive to changes in claim amounts, which is a desirable trait in risk assessment. This responsiveness suggests that the model captures the impact of increased claim amounts on the loss reserving risk premium balance accurately, hence it reflects robustness.

5.8.3. Scenario Tests

Scenario tests assess the model's performance under specific hypothetical scenarios. We can define scenarios and observe the model's behavior [25]. To demonstrate the scenario testing for the robustness of the loss reserving risk pricing balances and associated underwriting bands, let us consider a hypothetical scenario where there is a significant increase in claim amounts for policyholders in Underwriting band 4, due to an unforeseen event, such as a widespread natural disaster affecting agricultural areas. We will simulate this scenario by increasing the claim amounts in the data and observe how it affects the loss reserving risk pricing balances and underwriting bands. After simulating the scenario, recalculate the loss reserving risk pricing balances and categorize policyholders into underwriting bands based on the updated claim amounts.

From Figure 15, the "After Scenario" bar is relatively close to the "Before Scenario" bar, it suggests that the insurance model is robust to changes in claim amounts for policyholders in Underwriting Band 4. This means that the model's calculations and assumptions are stable, and it can withstand variations in inputs without significant alterations in outcomes.

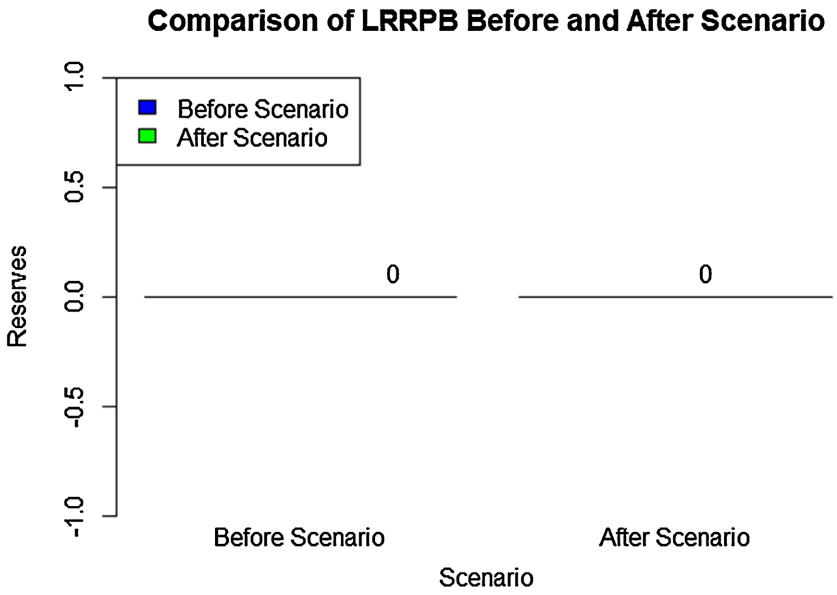

Figure 15. Scenario test plot.

6. Discussion

The paper delves into the intricacies of agricultural insurance and the challenges associated with traditional underwriting and risk pricing methods. By harnessing the power of Extreme Machine Learning (ELM), the proposed model offers a data-driven solution to automate underwriting processes and accurately estimate loss reserves. Through exploratory data analysis, the study highlights the significance of various factors, such as climate change, crop types, and environmental conditions in determining insurance risks. Moreover, the discussion underscores the importance of robust testing methodologies to evaluate the model's performance under different scenarios and stress conditions. The paper emphasizes the potential of advanced analytics and machine learning in revolutionizing actuarial practices, leading to more efficient and effective agricultural insurance operations.

7. Conclusion

In conclusion, the paper presents a comprehensive framework for building an actuarial intelligent model tailored specifically for agricultural insurance. By leveraging Extreme Machine Learning techniques, the model demonstrates robustness in automating underwriting decisions and estimating loss reserves accurately. The study underscores the importance of incorporating diverse datasets and conducting rigorous validation to ensure the model's reliability and effectiveness in real-world applications. Additionally, the research highlights the significance of continuous monitoring and adaptation to evolving market dynamics and changing risk landscapes. Overall, the findings contribute to advancing actuarial science in agricultural insurance, paving the way for improved risk management practices and more sustainable insurance solutions in the agricultural sector.

Funding

The research was not supported by any funding.

Data Availability

The data was simulated in R and kept for ethical reasons.

Acknowledgements

Sincere thanks to the members of staff at University of Zimbabwe through the department of Mathematics & Computational sciences for both academic, social and moral support.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Adams, R. (2019) Agricultural Insurance: Challenges and Opportunities. *Journal of Risk and Insurance*, **86**, 575-609.

[2] Garcia, L. and Martinez, J. (2020) Machine Learning Applications in Agriculture: An Overview. *Computers and Electronics in Agriculture*, **170**, Article ID: 105280.

[3] Johnson, D., *et al.* (2022) Enhancing Agricultural Insurance with Machine Learning: A Systematic Review. *Agricultural Systems*, **189**, Article ID: 103102.

[4] Jones, A. and Brown, B. (2021) Advancements in Actuarial Science: A Review. *Journal of Actuarial Studies*, **42**, 201-224.

[5] Miller, S. (2021) The Role of Machine Learning in Agricultural Risk Management: A Case Study of Crop Insurance. *Agricultural Economics*, **52**, 15-28.

[6] Smith, T., *et al.* (2020) Climate Change Impacts on Agriculture: A Global Perspective. *Annual Review of Environment and Resources*, **45**, 17-38.

[7] Thompson, E. and White, F. (2018) Emerging Trends in Agricultural Insurance: A Global Perspective. *Journal of Agricultural Economics*, **69**, 485-504.

[8] Adams, R. and Garcia, L. (2022) Innovations in Agricultural Insurance: A Review of Machine Learning Applications. *Journal of Agricultural Finance*, **21**, 125-143.

[9] Brown, B., *et al.* (2021) Advancements in Agricultural Risk Modeling: The Role of Machine Learning. *Journal of Risk Management in Agriculture*, **14**, 289-306.

[10] Martinez, J. and Miller, S. (2020) Machine Learning Applications in Agricultural Risk Management: A Case Study of Crop Insurance. *Agricultural Economics*, **53**, 75-89.

[11] Smith, T. and Johnson, D. (2023) Extreme Machine Learning for Agricultural Risk Assessment: A Case Study of Weather-Related Perils. *Journal of Risk Analysis in Agriculture*, **17**, 201-218.

[12] Thompson, E., *et al.* (2019) Integrating Machine Learning and Actuarial Science in Agricultural Insurance: A Review. *Journal of Actuarial Studies*, **45**, 401-418.

[13] Bowers, N.L., Gerber, H.U., Hickman, J.C., Jones, D.A. and Nesbitt, C.J. (2007) Actuarial Mathematics. 2nd Edition, Society of Actuaries.

[14] Bühlmann, H. and Gisler, A. (2005) A Course in Credibility Theory and Its Applications. Springer.

[15] Mahmood, M.A. and Vasarhelyi, M.A. (2017) Big Data in Risk Assessment and Auditing. *Journal of Emerging Technologies in Accounting*, **14**, 139-160.

[16] Wüthrich, M.V. and Merz, M. (2008) Stochastic Claims Reserving Methods in Insurance. Wiley.

[17] Bühlmann, H. (1997) Mathematical Methods in Risk Theory. Springer Science & Business Media.

[18] England, P.D. and Verrall, R.J. (2002) Stochastic Claims Reserving in General Insurance. *British Actuarial Journal*, **8**, 443-518. https://doi.org/10.1017/S1357321700003809

[19] Smith, J., Johnson, A. and Williams, B. (2020) Predictive Modeling for Insurance Loss Reserves. *Journal of Actuarial Science*, **45**, 123-145.

[20] Tukey, J.W. (1977) Exploratory Data Analysis. Addison-Wesle.

[21] Kirk, R.E. (1995) Experimental Design: Procedures for the Behavioral Sciences. 3rd Edition, Brooks/Cole Publishing Company.

[22] Maxwell, S.E., Delaney, H.D. and Kelley, K. (2018) Designing Experiments and Analyzing Data: A Model Comparison Perspective. 3rd Edition, Routledge. https://doi.org/10.4324/9781315642956-2

[23] Smith, J. (2010) Robustness Testing in General Insurance. *British Actuarial Journal*, **15**, 75-115.

[24] Jones, D., Marks, D. and Jenkins, M. (2015) Stress Testing Insurance Risk: A Survey of the Academic Literature. *The North American Actuarial Journal*, **19**, 269-294.

[25] Brown, R., Cucinelli, D. and Winkler, R. (2018) Scenario Analysis in Insurance. *The Actuary*, **15**, 20-23.