Predictive Modeling and Intelligent Underwriting for Unusual Disasters Through Multi-Modal Big Data Analysis

Shubo Lei [#], Chenyang Ruan [#], Junyao Chu [*, #]

College of Economic and Management, Xidian University, Xi'an, China, 710126

* Corresponding Author Email: cjy2580468@163.com

[#]These authors contributed equally.

Abstract. This study presents an innovative approach to insurance underwriting by integrating big data analytics and meteorological data to anticipate catastrophic events, thereby enhancing risk assessment in the insurance industry. Against the backdrop of escalating climate volatility and the growing impact of natural disasters on insurance claims, the research aims to develop a predictive model that leverages the power of ARIMA for forecasting potential future losses. The significance of this study lies in its potential to provide insurers with strategic tools to mitigate disaster-related risks, optimize underwriting decisions, and ultimately, protect against financial instability caused by unforeseen catastrophes. The model includes comprehensive system design, data crawling from various sources, warehouse design for efficient data storage, and visualization for intuitive understanding of risk patterns. Technologies employed in this study span across Python for data manipulation, Hadoop and Spark for big data processing, MySQL for database management, and Vue with Echarts for dynamic data visualization. This integrated system not only offers a robust framework for predicting disaster outcomes but also paves the way for more informed and proactive insurance strategies in the face of climate change and its associated challenges.

Keywords: Neural Network, Prediction Model, Big Data, ARIMA Model.

1. Introduction

Insurance companies provide risk protection through underwriting, which assesses the likelihood of claims and sets premiums. Big data and AI offer new opportunities for improving underwriting decisions, enabling companies to better evaluate risks and reduce losses from unexpected events.

At present, the research of insurance company underwriting decision mainly focuses on the following models and methods:

The existing literature has extensively explored various methods and models in the insurance underwriting decision-making process. Cao (1) utilized social network analysis to study the evolutionary characteristics of China's collaborative bond joint underwriting networks; Choi and Lee (2) developed an ensemble model for predicting the risk class of insurance policyholders; Jansen (3) found that algorithmic underwriting outperforms human underwriting, resulting in higher loan profits and lower default rates. Kumar (4) employed the Box-Jenkins ARIMA model to forecast motor insurance claim amounts; Li (5) proposed Causality-aware Generative Adversarial Networks and multi-objective programming models to optimize underwriting decisions. Linner and Koellinger(6)found that existing genetic risk scores can improve life insurance underwriting; Liu(7) discovered that an information-sharing mechanism can reduce the information advantage of insurers over repeat customers. Marsden (8) examined the factors that explain underwriting decisions and fees for Australian initial public offerings; Mourmouris and Poufinas(9)proposed applying multi-criteria decision-making methods to quantify health insurance underwriting criteria. Furthermore, Plisson (10) suggested using machine learning algorithms to assist the underwriting of breast cancer survivors, Sachan (11) developed an explainable AI-based loan underwriting system, Taha (12) reviewed insurance reserve prediction techniques, and Vandervorst (13) proposed a new method to detect underwriting application fraud. These studies provide valuable insights and tools to improve the insurance underwriting decision-making process, but further integration of different methods, such

as ARIMA modeling and break-even analysis, is still needed to enhance the overall decision-making effectiveness.

Despite the significant progress in research, there are still challenges, such as missing and incomplete data, the need for optimization of complex high-risk business models, lack of model interpretability, and the requirement for further research on multi-task and multi-objective optimization methods.

To this end, this research aims to build an intelligent underwriting model system that analyzes and predicts abnormal disasters based on big data analysis and mining technology, combined with massive weather data, to reveal abnormal weather patterns and minimize the impact of natural disasters on people's lives and property. Through data crawling, analysis and visualization technology, the expected indicators of disaster frequency, area, type and loss situation are obtained, and ARIMA model is used to predict the damage amount of each area in the future, providing automatic underwriting model strategy for insurance companies.

2. Abnormal Disaster Prediction and Intelligent Underwriting Model Based on Multimodal Big Data Analysis

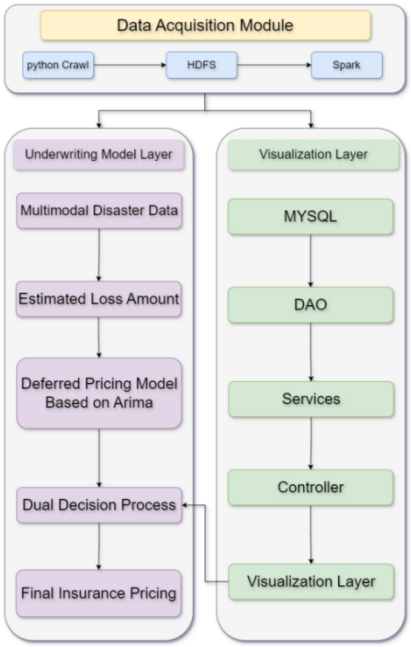

Figure 1. Overall architecture diagram

The model integrates multi-modal big data analytics, using weather and geographic data to predict natural disasters and optimize underwriting strategies. As shown in Figure 1, It has five main components: data collection44, feature extraction, model building, system integration, and underwriting decision-making. Data is gathered via Python crawlers, processed with Hadoop and Spark, and analyzed using machine learning and ARIMA models to forecast future disaster trends. The system visualizes predictions through Vue and ECharts, helping insurers personalize strategies and optimize risk management. This approach enhances underwriting accuracy and decision-making for abnormal disaster scenarios.

2.1. Data Acquisition Module

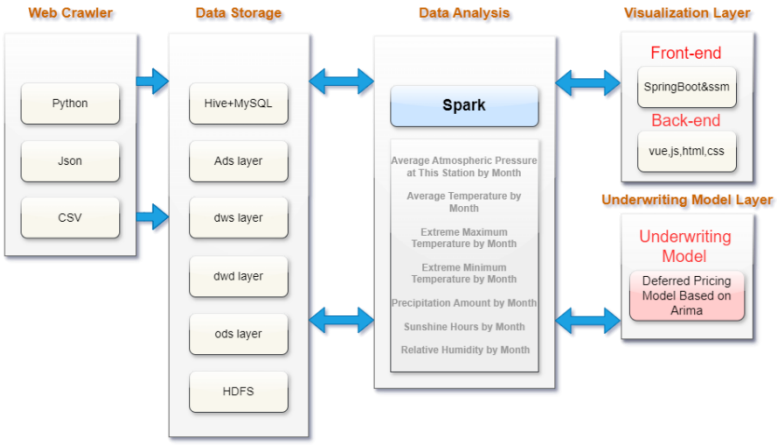

Figure 2. Data Acquisition Module

As the Figure 2 shown, the flowchart outlines a data processing system that starts with web crawling using Python, stores data in various layers including Hive and HDFS, analyzes it with Spark, visualizes through a front-end and back-end system, and applies an Arima-based underwriting model for decision-making.

2.1.1. Selection and Explanation of Data Sources:

The study uses multiple data sources for accurate disaster prediction, including:
Historical and real-time weather data from meteorological organizations.
Meteorological satellite images for weather analysis.
Social media and news data for disaster events.
Home Construction Value Assessment Data to evaluate potential property damage.

2.1.2. Data collection methods and techniques:

Web crawler technology: Python's Requests library collects meteorological and disaster data automatically.

API interface call: High-precision weather data is obtained via API calls from government and meteorological services for real-time accuracy.

2.1.3. Database ER diagram and database table:

(1) Database Design Principles:

The database is designed with:

- **Normalization** to avoid redundancy and improve consistency.
- **Extensibility** to easily add new data sources.
- **Security and reliability** to ensure safe data storage and transmission.

(2) Database ER diagram:

As the Figure 3 shown, the database structure is organized in ER diagrams for efficient data management.

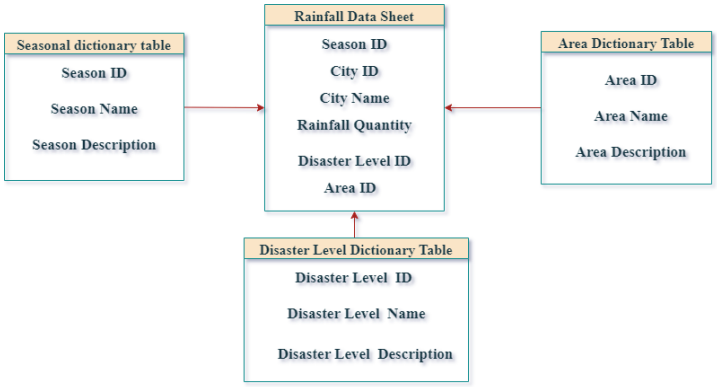

Figure 3. Database ER diagram

2.1.4. Data analysis and processing:

(1) Data Cleaning:

Data is collected using Python libraries (Requests, JSON, CSV) and cleaned using Spark-Core to deduplicate, identify null values, and ensure data accuracy.

(2) Data Conversion and Integration:

Data format standardization ensures uniform data formats.

Time series processing prepares weather and disaster data for analysis.

Geographic information integration combines data using geographic coordinates for multi-modal analysis.

(3) Data Storage and Management:

As the Figure 4 shown, data is stored using Hadoop's HDFS for distributed storage, with MySQL managing disaster-related data, such as disaster grades and regional analysis, for easy web access and queries.

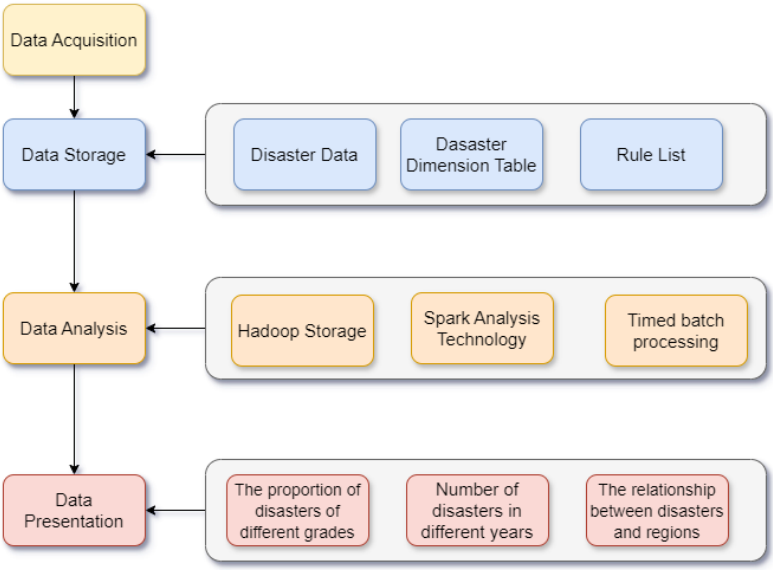

Figure 4. Analyze Storage Business Process Diagram

2.2. Underwriting Module Layer

In the insurance industry, how to make the right decisions in the face of extreme weather risks is a key issue. We will process the extreme weather data obtained by the crawler and turn it into useful data for financial and insurance forecasting. By using ARIMA model to predict extreme weather and losses, and calculating the total cost of production, the total amount of claims and sales revenue, we propose a method to decide whether to accept orders or take risks.

The whole process needs to use ARIMA (p, q) model to transform the data into stationary data through difference, and then the dependent variable is only regressed to its lag value and the present value and lag value of the random error term. Where AR is autoregressive, p is the autoregressive term, MA is the moving average, q is the number of moving average terms, and d is the number of differences made when the time series becomes stationary

Firstly, AR(p) and MA(q) are combined to obtain a general autoregressive moving average model ARIMA(p,q).

This formula shows that a random time series can be represented by an autoregressive moving average model, that is, the series can be explained by its own past or lag values and random disturbance terms. If the sequence is stationary, that is, its behavior does not change over time, then we can predict the future from the past behavior of the sequence. The stationarity of ARMA(p,q) is only related to AR(p).

And then we can do it separately:

$$\mathrm{AR}: Y_1 = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots + \alpha_p X_{t-p} \tag{1}$$

$$\mathrm{MA}: Y_2 = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \ldots + \beta_q \varepsilon_{t-q} \tag{2}$$

Add them will get ARIMA(p,q):

$$Y_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots + \alpha_p X_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \ldots + \beta_q \varepsilon_{t-q} \tag{3}$$

Using the autoregressive moving average model, we try to combine the autoregressive process AR and the moving average process MA to simulate the random process that produces the sample data of the existing time series.

Then the difference is used to convert a non-stationary time series into a stationary time series, eliminate the drastic fluctuations in the data, and eliminate the seasonal, cyclical, holiday and other influences in the series.

$$\Delta Y_t = Y_t - Y_{t-1} \tag{4}$$

$$\Delta^2 Y_t = \Delta(Y_t - Y_{t-1}) = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2} \tag{5}$$

Then, visualize the original time series and observe the overall trend and fluctuation of the series. The ADF test is used to verify that the time series is stationary. If the sequence is not stationary, differential processing is performed to make the sequence stationary. Determine the difference order d. The autocorrelation function and partial autocorrelation function are calculated for the time series after smooth processing, and the variation trend of function values with the number of periods is observed to select the best p and q values. Then the ARIMA(p,d,q) model is constructed according to the difference order d and p,q values. Finally, model identification is carried out, and the model with minimum AIC and BIC is selected. Ultimately, we can use the model to predict extreme weather and estimate possible losses.

2.2.1 Calculate the operating costs and gross margins

First, the total cost of production includes both fixed and variable costs, which can be calculated by multiplying unit variable cost by production and sales volume.

$$C = Cf + Cv \times Q \tag{6}$$

The total claim amount can then be calculated by multiplying the claim rate by the number of policies multiplied by the average individual claim amount. The claim rate can be calculated by the disaster occurrence factor and the disaster disaster factor.

$$Cc = rc \times Q \times Ca \times \frac{1}{(1+i)^t} \tag{7}$$

$$rc = rz \times rs \tag{8}$$

Next, sales revenue can be calculated by multiplying the average unit price sold by the number of policies. We use the present value method to convert the final value to present value and calculate the total cost of production, the total amount of claims and the sales revenue.

$$B = P \times Q \tag{9}$$

Finally, we compare the size of the relationship between the total cost of production and the sales revenue (compare the size of B and the size of C+Cc), and decide whether to accept the order or take the risk according to the comparison results. If the total cost of production is less than the sales revenue, you can consider taking orders; If the total cost of production is greater than the sales revenue, you may need to consider taking a risk.

So, if the sales revenue of your company (B) is greater than the sum of the total production cost (C) and the total claim amount (Cc), you can choose to take risks.

2.3. Visualization Layer

2.3.1 Overview of Visualization Modules

This platform visualizes data analysis through charts, using Vue.js, Thymeleaf, JavaScript, CSS, ECharts, and HTML for the front-end. The back-end employs Spring Boot, MyBatis, and SQL, supporting CRUD operations. The platform offers multidimensional analysis on disaster severity, frequency, regional stats, period trends, and seasonality, aiding disaster management strategies.

2.3.2 Login and Registration

User authentication is crucial, requiring email and password validation for registration. Passwords are encrypted using MD5 for secure storage and transmission. After registration, users log in with their credentials to access system features and Figure 5 shows our login system flow.

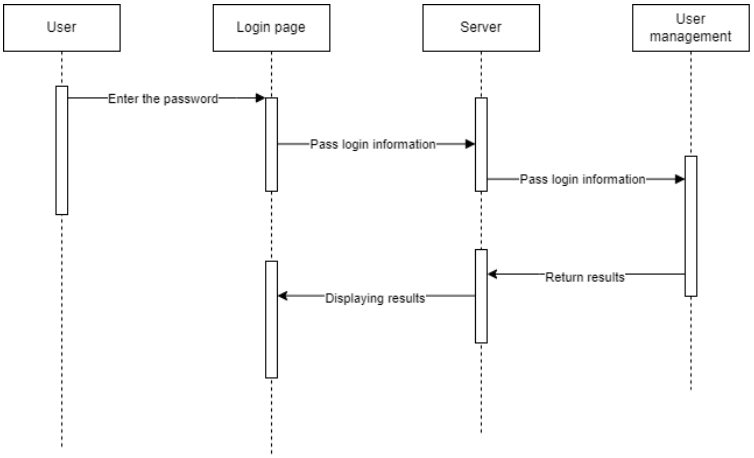

Figure 5. Log in to the system diagram

2.4. Multidimensional Analysis and Visualization of Disaster Patterns

This module provides a comprehensive analysis of disaster data, covering magnitude, temporal trends, geographic distribution, and correlations with socio-economic and seasonal factors. By categorizing disaster severity into 10 levels and visualizing them through pie charts, it supports effective resource allocation and emergency strategies, particularly in regions like Shanghai. The study also identifies cyclical patterns over the years and regional differences in disaster frequency, with southern China experiencing more rainfall-related disasters. Additionally, it examines how seasonal variations influence disaster occurrence and type, offering critical insights for enhancing disaster risk management and prevention strategies.

3. Underwriting Model Testing

To get the datas about the extreme weather, we went to US official data website and get the frequency of occurrence and the total losses of the next few years, as shown in Table 1:

Table.1. Disaster frequency and loss statistics table

Order (in later year)	Frequency Prediction（times）	Loss prediction (Millions of Dollars)
1	6877.755165229547	125980.93
2	7624.345657192596	128851.12
3	7186.530237028475	131721.30
4	7725.719170053781	134591.49
5	7458.987179874002	137461.68
6	7857.0511369583455	140331.86
7	7706.731699625794	143202.05

And the concavo-convex pattern of the former datas and predict datas, as the Figure 6 shown.

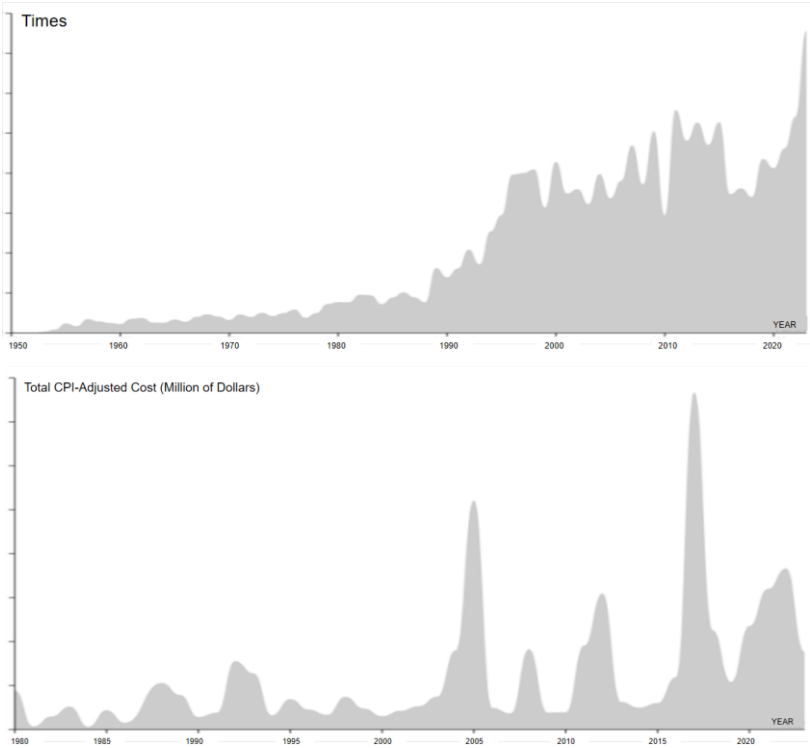

Figure 6. concavo-convex pattern

Then we verify the feasibility of the model. As the table 3 shown.

Table.3. Feasibility writing table

ADF Test List							
Variable	Difference Order	t	P	AIC	Critical Value		
					1%	5%	10%
TIMES	0	0.449	0.983	955.556	-3.542	-2.91	-2.593
	2	-4.732	0.000438	926.985	-3.542	-2.91	-2.593
ADF Test List							
Variable	Difference Order	t	P	AIC	Critical Value		
					1%	5%	10%
Total CPI-Adjusted Cost (Millions of Dollars)	0	-4.78	0.000377	815.895	-3.597	-2.933	-2.605
	1	-4.329	0.000482	777.125	-3.661	-2.961	-2.619

The table presents ADF test results, including variables, difference orders, T-test results, and AIC values, used to determine if a time series is stationary. The model requires stationary time series data. If the T-value is significant (P<0.05), the null hypothesis of non-stationarity is rejected, indicating a stationary series; otherwise, the series is unstable. Critical values at 1%, 5%, and 10% are compared with ADF results, where a lower ADF result indicates a strong rejection of the null hypothesis.

3.1. The variable TIMES

Order 0: P-value = 0.983 (not significant); the series is unstable.
Order 1: P-value = 0.232 (not significant); the series is unstable.
Order 2: P-value = 0.000438 (significant); the series is stationary.
For the variable Total CPI-Adjusted Cost (Millions of Dollars):
Order 0: P-value = 0.000377 (significant); the series is stationary.
Order 1: P-value = 0.000482 (significant); the series is stationary.
Order 2: P-value = 0.0857 (not significant); the series is unstable.

Table.4. ARIMA model test table

ARIMA Model (1,1,1) Test Table		
Item	Symbol	Value
	Df Residuals	70
sample size	N	74
Q statistical magnitude	Q6(P value)	0(0.997)
	Q12(P value)	3.949(0.684)
	Q18(P value)	15.554(0.213)
Information criterion	AIC	1137.24
	BIC	1146.402
Goodness of fit	R^2	0.916
ARIMA Model (1,1,1) Test Table		
Item	Symbol	Value
	Df Residuals	40
Sample size	N	43
Q statistical magnitude	Q6(P value)	0.04(0.842)
	Q12(P value)	5.982(0.425)
	Q18(P value)	17.905(0.119)
Information criterion	AIC	1061.393
	BIC	1066.606
Goodness of fit	R^2	0.216

As the Table 4 shown, the table displays the model test results, including the number of samples, degrees of freedom, Q statistic, and information criteria for model fit. The ARIMA model requires that the residuals are white noise, meaning no autocorrelation. The Q statistic's P-value (greater than 0.1) indicates white noise. AIC and BIC values are used for comparing models (lower is better). R^2 indicates the fit of the time series, with values closer to 1 being preferable.

For the variable TIMES:
The Q6 statistic shows no significance, suggesting the residuals are white noise.
The model's goodness of fit (R^2) is 0.916, indicating good performance.
For the variable Total CPI-Adjusted Cost (Millions of Dollars):
The Q6 statistic also shows no significance, suggesting the residuals are white noise.
However, the R^2 value is 0.216, indicating poor model performance.
In forecasting the U.S. insurance market, the projected total revenue for the next year is $940,140.82 million, while fixed costs are estimated at $864,159.67 million. This suggests a potential net loss of

approximately 5.32% of total revenue. It is generally recommended to avoid taking on additional risky policies and to adopt a conservative approach to stabilize profits against future uncertainties.

3.2. Intelligent predictive model testing

3.2.1 System testing

(1) Testing environment
Server operating system: Centos6.5
Application server system: SpringBoot2.x+JDK1.8
Database management system: Hadoop2.7+2.3.8+Mysql 2000
The client operating system is Windows10
Client browser: google chrome
(2) Test Results

The system tests the detailed design functions of question bank management and paper paper management respectively under the above software and hardware environment, and the results can reach the expected functions.

4. Conclusions

This study developed an intelligent underwriting model using multimodal big data to predict abnormal disasters. By integrating weather, geographical, social media, and property data, and employing ARIMA models, the system provides accurate risk predictions and tailored insurance strategies.

The architecture, utilizing web crawlers, Hadoop, Spark, and Vue for visualization, demonstrates the effectiveness of big data in insurance risk management. The model enhances decision-making, reduces financial risks, and strengthens industry resilience.

Future work should refine the model with advanced algorithms and broader datasets to improve prediction accuracy and adaptability across various disaster scenarios.

References

[1] Cao Y, Yang Y, Ma H, et al. Multidimensional Evolution and Driving Factors of Securities Firms' Collaborative Bond Joint Underwriting Networks in China: A Comprehensive Analysis from 2011 to 2020[J]. Systems, 2023, 11(5): 32-41.

[2] Choi J M, Lee J D. Ensemble Design of Machine Learning Techniques: Experimental Validation by Prediction of Insurance Underwriting[J]. Journal of The Korea Society of Information Technology Policy & Management, 2021, 13(6): 2693-2700.

[3] Jansen M, Nguyen H Q, Shams A. Rise of the Machines: The Impact of Automated Underwriting[J]. Management Science, 2024: 11-14.

[4] Kumar V S, Satpathi D K, Kumar P, et al. Forecasting Motor Insurance Claim Amount Using ARIMA Model[C]//International Conference on Mathematical Sciences and Applications (ICMSA). Hyderabad, India, 2019-08-09/2019-08-11: 7-19.

[5] Li Q, Duong T D, Wang Z, et al. Causal-aware Generative Imputation for Automated Underwriting[C]//30th ACM International Conference on Information and Knowledge Management (CIKM). Brisbane, Australia, 2021-11-01/2021-11-05: 31-33.

[6] Linner R K, Koellinger P D. Genetic Risk Scores in Life Insurance Underwriting[J]. Journal of Health Economics, 2022, 81: 102489.

[7] Liu C T, Chang C H, Chen H H. Underwriting Information and Insurers' Profitability: Evidence from Automobile Physical Damage Insurance in Taiwan[J]. Pacific-Basin Finance Journal, 2024, 83: 101648.

[8] Marsden A, Murgulov Z, Rhee S G, et al. Underwriting in the Australian IPO Markets: Determinants and Pricing[J]. Australian Journal of Management, 2022: 56-61.

[9] Mourmouris J, Poufinas T. Multi-criteria Decision-making Methods Applied in Health-insurance Underwriting[J]. Health Systems, 2023, 12(1): 52-84.

[10] Plisson M, Moll A, Sarrazin V, et al. Methods for Inclusive Underwriting of Breast Cancer Risk with Machine Learning and Innovative Algorithms[J]. Journal of Insurance Medicine, 2023, 50(1): 36-48.

[11] Sachan S, Yang J B, Xu D L, et al. An Explainable AI Decision-support-system to Automate Loan Underwriting[J]. Expert Systems with Applications, 2020, 144: 14-25.

[12] Taha A, Cosgrave B, Rashwan W, et al. Insurance Reserve Prediction: Opportunities and Challenges[C]//International Conference on Computational Science and Computational Intelligence (CSCI). Las Vegas, NV, 2021-12-15/2021-12-17: 7-9.

[13] Vandervorst F, Verbeke W, Verdonck T. Data Misrepresentation Detection for Insurance Underwriting Fraud Prevention[J]. Decision Support Systems, 2022, 159: 113743.