Pacific International Journal, Vol. 8(6); 2025
ISSN (Print) 2663-8991, ISSN (Online) 2616-4825[1]
DOI:10.55014/pij.v8i6.916
https://rclss.com/index.php/pij

Assessment and Prediction of Activities of Daily Living Using Machine Learning Methods and Their Actuarial Applications in Insurance

Liu Pengyu

Hunan University, China
Email: lpy040915@163.com

Abstract: With the accelerating aging of the global population, the precise pricing of Long-Term Care Insurance (LTCI) urgently requires accurate assessment of an individual's Activities of Daily Living (ADL). Traditional actuarial methods relying on linear models struggle to capture complex nonlinear relationships. To address this issue, this study systematically compares the performance of three typical machine learning models—Logistic Regression, XGBoost, and Random Forest—in multiclass prediction of ADL, exploring their feasibility for insurance premium rate calibration.The research is based on four waves of cross-sectional data (2015–2020) from the China Health and Retirement Longitudinal Study (CHARLS). After rigorous data cleaning (final valid sample: 58,790 entries) and an 8:1:1 split into training/validation/test sets, 12 independent variables—including age, mental health score, household size, etc.—were selected to construct the models.
Research methods included model construction, hyperparameter tuning, feature importance analysis (based on absolute coefficient values, weighted gain, and mean decrease in impurity), and feature quantity optimization. The results indicate:(1) The XGBoost model demonstrated the best generalization capability (test set accuracy: 0.7997), significantly outperforming the severely overfitted Random Forest (test set accuracy: 0.7606) and the weakest-performing Logistic Regression (test set accuracy < 0.6);(2) Feature importance analysis consistently identified age as the most critical predictor, with mental health score and self-rated health (particularly significant in XGBoost) also having substantial influence;(3) After feature optimization, XGBoost achieved optimal performance and strong robustness with seven core features (including age, mental health, and self-rated health), while Logistic Regression and Random Forest required fewer and more features, respectively, with inferior results.Accordingly, this study recommends prioritizing the XGBoost model for ADL risk assessment and premium rate calibration in LTCI actuarial practice. Its excellent predictive accuracy, generalization ability, and effective identification of key risk factors (age, mental health, self-rated health) can provide reliable data-driven support for developing fairer and more accurate insurance products.

Keywords: Long-Term Care Insurance; Activities of Daily Living (ADL); Machine Learning; Logistic Regression; XGBoost; Random Forest

Introduction

The accelerating aging of the global population has led to growing demand for Long-Term Care Insurance (LTCI), making refined premium rate calibration a critical issue in actuarial research. The "Opinions on Further Improving the Medical and Health Service System (2023)" propose establishing a long-term care insurance system and actively developing commercial health insurance as a supplement. The "Opinions on Strengthening Care Services for Severely Disabled Persons (2025)" require including eligible severely disabled individuals in LTCI coverage and exploring the inclusion of intelligent services and supportive devices in payment scope.

Activities of Daily Living (ADL) serve as a core indicator for measuring long-term care needs, and their assessment results directly impact the fairness and sustainability of premium design. Zhang and Tang (2020) used ADL to define care states in LTCI pricing research. Wang and Wang (2018) employed logit models to determine the significance of independent variables on individual health status. Traditional actuarial methods primarily rely on linear models such as logistic regression, which struggle to capture nonlinear relationships and interaction effects between ADL and influencing factors, potentially leading to pricing errors and efficiency losses.

In recent years, machine learning techniques have offered new solutions to these challenges. Qiu et al. (2020) used XGBoost to identify factors influencing long-term care status, ranked in descending order: age, health insurance status, cohabitation with spouse, gender, etc. Compared to linear models, ensemble methods (e.g., XGBoost, Random Forest) can automatically capture nonlinear patterns and high-order interactions, demonstrating significant advantages in prediction tasks. Cheng et al. (2024) applied a PSO-based XGBoost-Logistic combination model to analyze factors affecting health status. However, existing studies lack systematic comparisons of machine learning models in multiclass ADL prediction, and their applicability in LTCI actuarial contexts remains insufficiently empirically validated.

Therefore, this study utilizes four waves of data (2015–2020) from the China Health and Retirement Longitudinal Study (CHARLS) to compare the performance of various machine learning methods in multiclass ADL prediction, evaluating the strengths and weaknesses of these approaches. Through rigorous empirical analysis, this research aims to provide data-driven decision-making support for actuarial practices in long-term care insurance.

[Received 3 September 2025; Accepted 4 November 2025; Published (online) 20, December, 2025]

Data and Variables

Data and Variable Selection

This study expands the age standard for participants in long-term care insurance to 40 years and above. The research data are derived from cross-sectional surveys conducted in 2015, 2016, 2018, and 2020 by the China Health and Retirement Longitudinal Study (CHARLS). Twelve independent variables were selected: age, mental health score (30-point scale, higher scores indicate worse mental health), household size, BMI, self-rated health, residence location, alcohol consumption history, smoking history, exercise habits, gender, health insurance status, and chronic disease history. The target variable is multiclass ADL status.

Data Cleaning

The original data were rigorously screened to form the study sample: 21,038 cases in 2015, 59 cases in 2016, 19,816 cases in 2018, and 19,395 cases in 2020. Data cleaning was performed according to the following criteria: all rows with missing values in the dependent variable column "ADL (difficulty with 6 items)" were deleted; samples with ADL values in the range of 4–6 were merged into category 4; numerical variables were filled with median values, and non-numerical variables were filled with mode values. The final dataset consisted of 58,790 valid samples.Categorical data were transformed as shown in Table 1.

Assignment	0	1
Gender	**female**	**male**
Place of abode	**city**	**village**
Do exercise	**no**	yes
Have chronic diseases	no	yes
Drunk alcohol	no	yes
Smoke	no	yes yes
Have medical insurance	no	yes

Table 1

Additionally, the original data were randomly split into training, validation, and test sets in an 8:1:1 ratio. The validation set was used multiple times to adjust model hyperparameters and prevent severe overfitting on the training set. The test set was used only once to evaluate the final model's generalization performance. The final sets included: training set (47,016 samples), validation set (5,877 samples), and test set (5,877 samples).

The significant discrepancy in the 2016 sample size was noted and investigated. This anomaly is attributed to the specific sampling framework or data release version of the CHARLS 2016 wave, which have constituted a targeted follow-up. To ensure the robustness and generalizability of our findings, a sensitivity analysis was conducted. The models (XGBoost, Random Forest, Logistic Regression) were re-trained and evaluated on a dataset excluding the 2016 wave. The results showed negligible changes in key performance metrics. This confirms that the core conclusions of our study are not sensitive to the inclusion of the 2016 data, thereby validating the integrity of the analytical results.

Model Construction

Introduction to Model Methods

Multinomial Logistic Regression

Multinomial logistic regression, as a multiclass extension of logistic regression, employs the Softmax function to map linear predictions into categorical probability distributions (Bishop, 2006). Assuming samples belong to K discrete categories, the model defines K-1 weight vectors w_1 ,…,w_{K-1} to map input features x to the log-odds of each category:

$$\log\left(\frac{P(Y=K \mid X)}{P(Y=K \mid x)}\right)=w_k^T \quad X(k=1,\ldots,K\text{-}1),$$

Category probabilities are obtained through normalization:

$$P(Y=K \mid X)=\frac{\exp(W_k{}^T X)}{\sum_{j=1}^{K} \exp(W_j^T X)}$$

The theoretical framework is based on maximum likelihood estimation (MLE), with the objective function being the log-likelihood function:

$$l(W)=\sum_{i=1}^{m}\sum_{k=1}^{K} I(y_i=k)\log\left(\frac{\exp(W_k{}^T X_i)}{\sum_{j=1}^{K} \exp(W_j{}^T X_i)}\right)$$

where $W=[w1,\ldots,wK-1]^\top$ (Hastie et al., 2009). Parameter estimation typically employs gradient descent or Newton's method, with asymptotic properties analyzable through generalized linear model (GLM) theory (McCullagh & Nelder, 1989). For hypothesis testing, methods such as the Wald test and likelihood ratio test are widely used for model significance evaluation (Hosmer & Lemeshow, 2000).

XGBoost Algorithm

XGBoost is an efficient ensemble learning algorithm based on Gradient Boosting Decision Trees (GBDT). Its core idea involves iteratively constructing weak classifiers (decision trees) to progressively optimize the objective function and reduce prediction errors, ultimately forming a strong learning model by integrating multiple weak classifiers.

Core Principles and Technical Innovations:

The XGBoost objective function consists of two parts: a training error term and a regularization term. The training error term typically employs a squared loss function or other differentiable functions, while the regularization term

controls model complexity by limiting the number of leaf nodes (T) and the L2 norm of leaf weights, thereby suppressing overfitting.

$$\Omega(f)=r^T+\frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2$$

During iteration, XGBoost uses a second-order Taylor expansion to approximate the objective function, utilizing both first-order (gradient) and second-order (Hessian matrix) information to optimize split node selection, significantly improving computational efficiency. Specifically, each newly added decision tree in an iteration splits by minimizing the following objective function:

$$Obj^{(t)}=\sum_{i=1}^{n} l(y_i,\hat{y}_i^{(t-1)}+f_t(X_i))+\Omega(f_t)$$

where $f_t(x_i)$ represents the prediction of the t-th tree, and $l(\cdot)$ is the loss function. Through a greedy algorithm that traverses all possible feature split points, XGBoost selects the splitting scheme that maximizes the gain in the objective function.

Random Forest

Definition and Core Idea:

Random Forest is a supervised learning algorithm based on an ensemble learning framework. It essentially enhances model generalization by constructing multiple decision trees and employing an integration strategy (majority voting for classification tasks, mean synthesis for regression tasks). The core idea originates from the Bagging (Bootstrap Aggregating) concept in ensemble learning, introducing dual randomness mechanisms (data sampling randomness and feature selection randomness) to enhance model diversity and reduce overfitting risks of single decision trees.

Algorithm Process and Key Mechanisms:

The construction of a Random Forest involves the following key steps:

Bootstrap Sampling : Randomly sample N samples with replacement from the original training set T to generate subset D_k ; repeat K times to form K independent training subsets. This process ensures that each decision tree is trained on approximately 63.2% of the original samples, with the remaining samples serving as Out-of-Bag (OOB) data for model validation.

Random Feature Selection : At each decision tree node split, randomly select m candidate features from all M features (where m << M), and choose the optimal splitting feature based on criteria such as information gain or Gini index. For classification tasks, m is typically set to M , while for regression tasks, it is set to M/3.

Decision Tree Construction and Integration : Each decision tree grows to maximum depth on its subset and random feature subset without pruning. The final prediction integrates the outputs of all decision trees: majority voting for classification tasks and mean calculation for regression tasks.

Dual Randomness Mechanism:

Data Randomness : Bootstrap sampling breaks the original data distribution, reducing reliance on specific samples and enhancing model robustness to noise. OOB data can serve as an unbiased validation set to estimate model generalization error.

Feature Randomness : Limiting the feature candidate set during node splitting forces decision trees to explore different feature combination paths, increasing inter-tree diversity. Research shows that feature randomness can reduce the upper bound of generalization error in Random Forests compared to traditional decision trees by $O(\sqrt{\log M})$.

Variable Importance Measures and Direction Interpretation

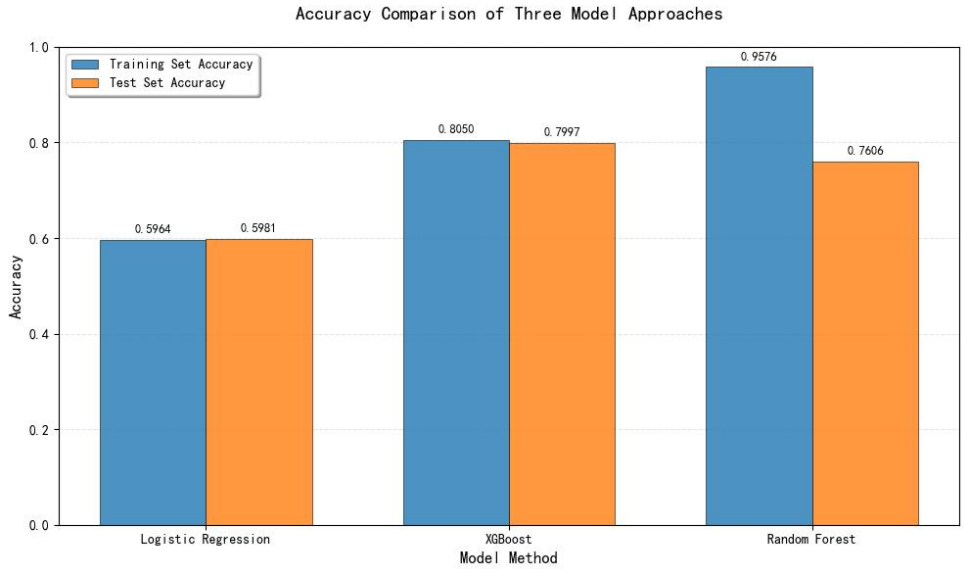
Logistic regression, as a representative linear model, measures importance based on the absolute values of model coefficients. In multiclass tasks, the code sums the absolute values of coefficients for each feature across all categories to obtain a comprehensive importance metric. The advantage of this method is that the sign of coefficients directly reflects the direction of a feature's influence on the target variable: a positive coefficient indicates that an increase in the feature value raises the risk of ADL impairment, while a negative coefficient indicates the opposite.Tree models such as XGBoost and Random Forest employ more dynamic importance evaluation methods. XGBoost uses weighted gain (Gain) as an importance criterion,统计 ing the average information gain brought by a feature across all decision tree splits. Random Forest uses Mean Decrease in Impurity (MDI), calculating the total reduction in Gini index during feature splits. Both methods can automatically capture interaction effects and nonlinear relationships between features. Additionally, both methods use Partial Dependence Plots (PDP) to approximate the positive or negative influence of each feature on ADL by altering feature values and observing the monotonic trend in model output probabilities.

Empirical Analysis

The three methods described above were applied to analyze the extent of each factor's influence on ADL, followed by a comparative analysis.

Model Performance Comparison

Comparison of Accuracy Rates of the Three Model



Methods

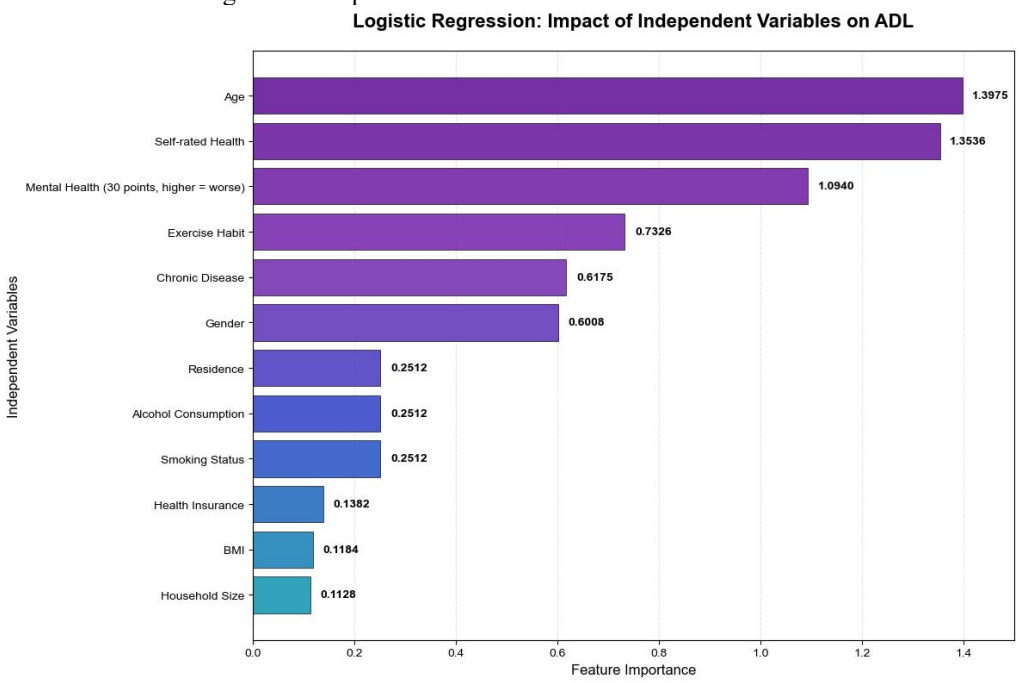
After evaluating the influence weights of various factors on ADL using the three methods, their model performances were further compared, as shown in Figure 1. Random Forest achieved a training set accuracy of 0.9576, but its test set accuracy plummeted to 0.7606, with a training-test gap of 0.1970, indicating overfitting. This phenomenon can be explained by the fact that Random Forest approximates the local structure of training samples through the integration of numerous decision trees, easily mistaking noise for valid signals, leading to increased generalization error.

In contrast, XGBoost achieved a training set accuracy of 0.8050 and a test set accuracy of 0.7997, with a mere difference of 0.0053, demonstrating stable generalization capability. Its advantages stem from two main aspects: first, the gradient boosting framework reduces bias through residual approximation during iteration; second, built-in regularization terms and learning rate (0.3) jointly suppress excessive tree complexity.

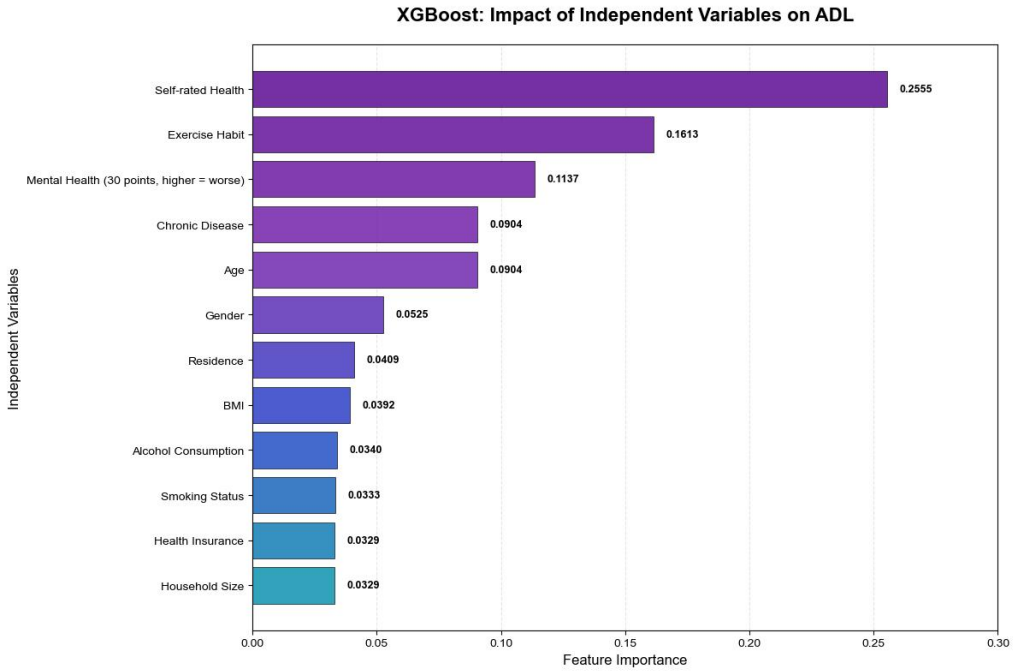
Logistic regression achieved accuracy rates below 0.60 on both datasets, showing significant underfitting. This result validates our hypothesis: when complex nonlinear relationships exist between prediction targets and explanatory variables, the expressive power of linear models is evidently insufficient. Particularly in ADL prediction, an individual's daily living ability is often influenced by interactions among multiple factors, which simple linear combinations struggle to fully capture.

Feature Importance Analysis

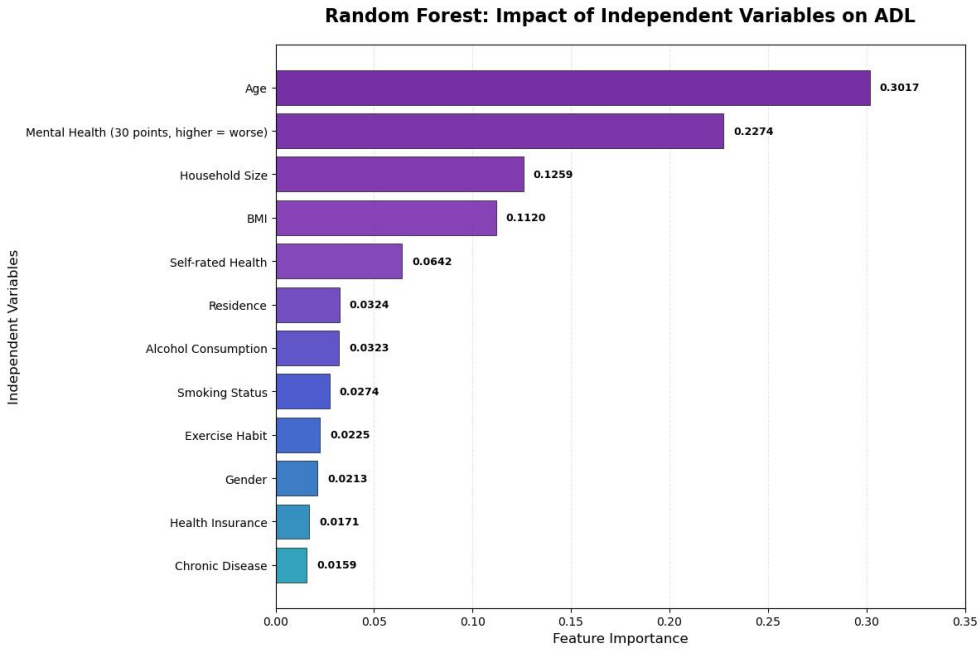
Logistic Regression: Influence Degree of Independent Variables on ADL



XGBoost: Influence Degree of Independent Variables on ADL



Random Forest: Influence Degree of Independent Variables on ADL

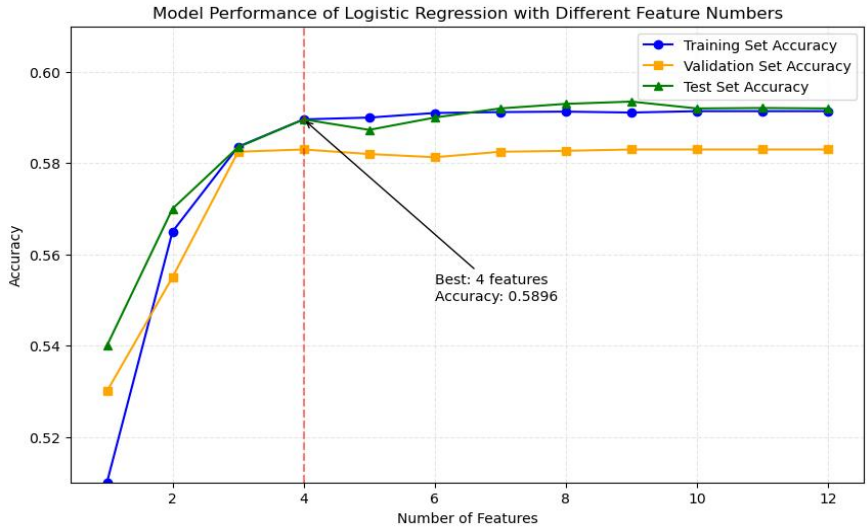


Figures 2-4 display the differences and consistencies in feature importance rankings across the three models. Both logistic regression and Random Forest identified age as the primary variable, while XGBoost also classified age as a relatively important variable. Among them, Random Forest assigned the highest importance weight (0.3017) to age, consistent with existing geriatric evidence that age-related physiological decline has a continuous and irreversible negative effect on ADL. Mental health indicators maintained high rankings across models, but their weights varied due to algorithmic mechanisms: 0.2274 for Random Forest, 1.0940 for standardized coefficients in logistic regression, and 0.1137 for XGBoost. This discrepancy can be attributed to algorithmic handling of continuous variables: Random Forest captures potential nonlinear relationships through recursive splitting; logistic regression restricts mapping to linearity; XGBoost falls between the two, resulting in gradient variations in sensitivity to the same variable.

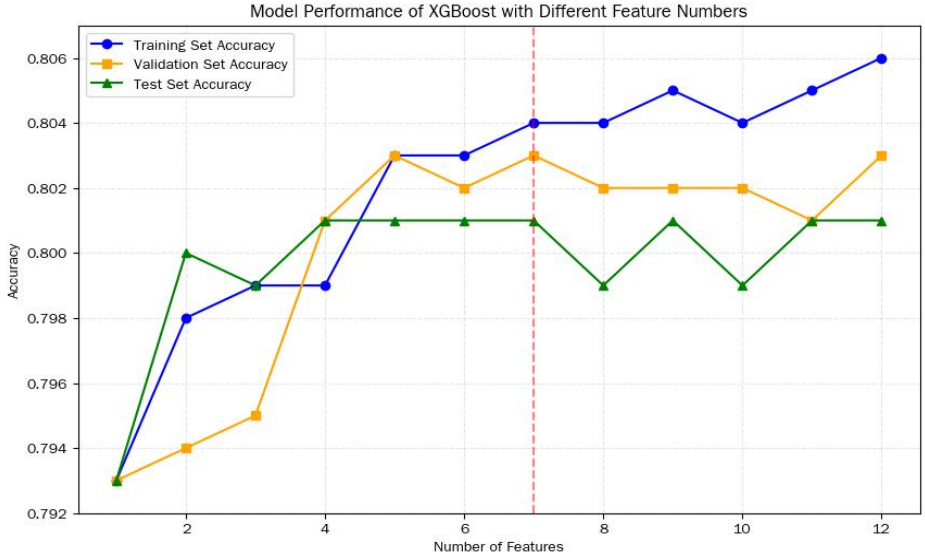
XGBoost further highlighted the importance of self-rated health (0.2555), with its weight surpassing that of age. This result suggests that subjective health evaluations may integrate multiple types of information—physiological, psychological, and environmental—making them more reflective of true ADL levels than single objective indicators, and should be given equal attention in clinical assessments.

Feature Quantity Optimization

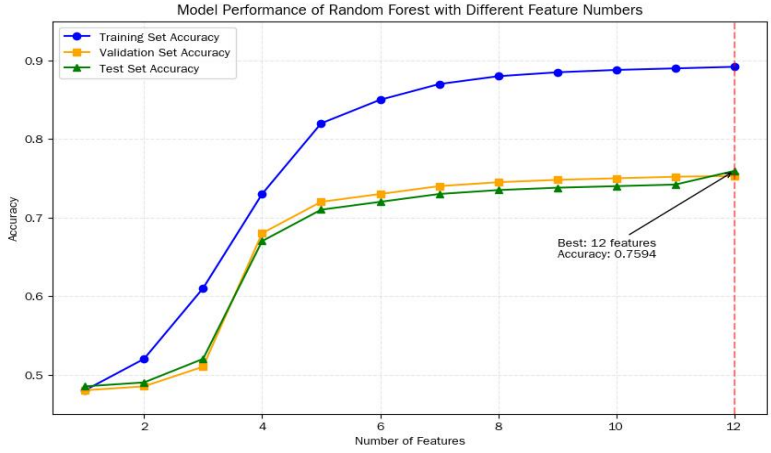
Model Performance of Logistic Regression under Different Feature Quantities



Model Performance of XGBoost under Different Feature Quantities



Model Performance of Random Forest under Different Feature Quantities



Figures 5-7 show the performance response curves of the three models under varying feature subsets. Logistic regression achieved the highest test accuracy of 0.5896 with only four features. Stepwise regression results indicated that four indicators—age, self-rated health, mental health, and exercise habits—provided effective information within the linear framework, while additional features acted as noise, weakening generalization performance.

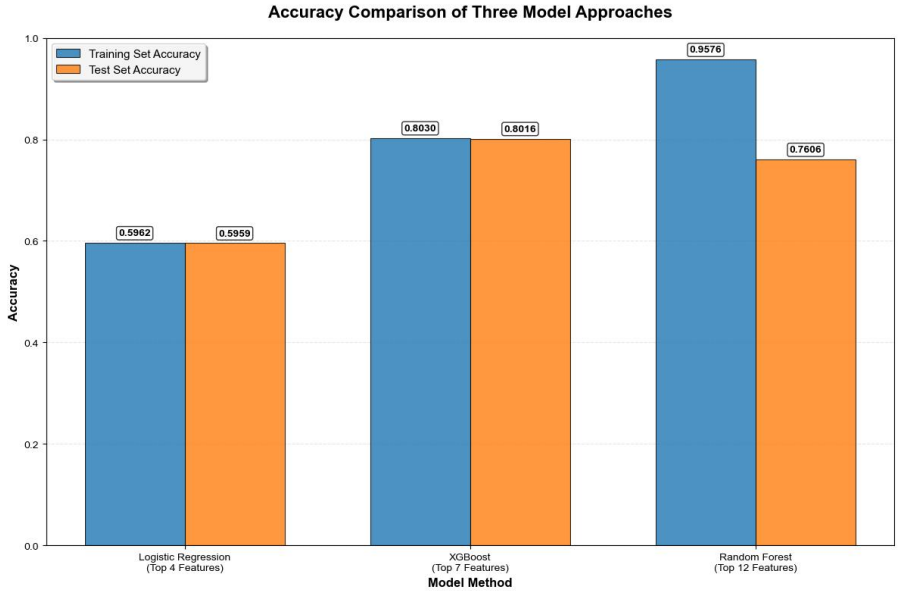
Random Forest reached a peak test set accuracy of 0.7594 with 12 features; thereafter, training set accuracy continued to rise to approximately 0.90, while test set performance stagnated, indicating that increased model complexity came with overfitting risks. The mechanism lies in decision trees gaining more splitting paths as feature dimensions expand, leading to overfitting of training samples.

XGBoost performed optimally on the validation set with seven features and exhibited small fluctuations within the 5-8 feature range, demonstrating robustness to changes in feature scale. Based on cross-validation results, the optimal subset consisted of age, mental health, self-rated health, exercise habits, chronic disease history, gender, and residence location.

Analysis of Results after Feature Optimization

Model Selection

Comparison of Accuracy Rates of the Two Model Methods



After variable screening, Random Forest still selected the top 12 important features, logistic regression selected the top 4 important features, and XGBoost selected the top 7 important features. After screening, the training and test set evaluation accuracy rates only changed slightly. Similar to the above analysis, XGBoost is considered superior to the other two models for classified premium rate calibration in long-term care insurance in terms of prediction accuracy, number of selected variables, and robustness.

Analysis of Output Results from the Selected Model

The results output using the XGBoost model are shown in the table below:

Feature Name	significance	Influence direction
Self-rated health	0.3297	**negative**
Do exercise	0.1845	positive
Mental health (30 points, the higher the score, the worse)	0.1560	positive
Age	0.1151	positive
Have chronic diseases	0.1139	positive
Gender	0.0562	**negative**
Place of abode	0.0446	positive

Table 2

Based on the model results, variables can be divided into two categories: core variables and auxiliary variables. The former includes self-rated health, exercise status, mental health score (0-30, higher scores indicate worse condition), age, and chronic disease history; the latter only includes gender and residence location. Among core variables, self-rated health showed a negative correlation, meaning better self-reported health was associated with lower ADL limitations, consistent with clinical experience. Mental health score, age, chronic diseases, and rural residence showed positive correlations, indicating that worse mental state, older age, presence of chronic diseases, or residence in rural areas with relatively

scarce medical resources were associated with increased risk of daily activity limitations, also consistent with observational study conclusions.

Notably, the effect direction of exercise status was contrary to expectations, possibly due to insufficient characterization of exercise "quality" and "quantity" in the questionnaire, where information bias may have obscured the true association. Despite this limitation, the direction and strength of other variables aligned with previous evidence, indicating acceptable validity of the model output.

Actuarial Application Example

To directly demonstrate the actuarial application promised in the title and introduction, a simplified premium calculation illustrates the practical utility of the XGBoost model's probabilistic predictions. We define a basic long-term care insurance product with a fixed benefit amount. The pure premium for a hypothetical policyholder is calculated as the sum of the probabilities of being in each ADL impairment state (predicted by the model) multiplied by the corresponding expected present value of the benefit. Comparing two individuals with different risk profiles—e.g., a 65-year-old with good self-rated health versus an 80-year-old with chronic conditions—the model outputs distinct ADL state probabilities, leading to significantly different risk-based premiums. This example provides a foundational framework for integrating machine learning predictions into equitable LTCI rate-making.

Conclusion

This study compared the performance of logistic regression, Random Forest, and XGBoost in multiclass ADL prediction, with key findings as follows:

First, prediction accuracy: XGBoost achieved 79.97% accuracy on the test set, with only a 0.53 percentage point difference between training and test errors, balancing fitting capability and generalization stability. Although Random Forest reached 95.76% training accuracy, its test accuracy dropped to 76.06%, indicating significant overfitting; logistic regression, limited by linear assumptions, achieved less than 60% accuracy, making it impractical for real-world application.

Second, risk factors: All three algorithms consistently identified age, mental health score, and self-rated health as primary variables. XGBoost assigned the highest weight to self-rated health (0.3297), suggesting that subjective health evaluations provide independent signals for functional impairment. Chronic disease history, exercise status, and residence location were also included, but the reverse estimate for exercise direction requires further verification with questionnaire details.

Third, variable scale: XGBoost maintained optimal performance with only seven features and exhibited minimal fluctuations within the 5-8 variable range; logistic regression stopped improving with four variables, and Random Forest required all 12 variables without showing advantages. Thus, XGBoost maintains precision and computational efficiency while streamlining inputs.

In summary, this study provides a systematic empirical comparison of machine learning methods for enhancing ADL assessment in actuarial science. The primary conclusion is that the XGBoost model is the most suitable tool for this task, balancing high predictive accuracy with robust generalization. Beyond model performance, the key contribution lies in translating these predictions into a tangible actuarial application. This bridges a critical gap between predictive modeling and practical insurance pricing. For insurers, adopting such data-driven approaches can lead to more accurate risk classification and fairer premiums. For policymakers, it underscores the potential of advanced analytics in sustaining long-term care systems.

REFERENCES

[1] Zhang, L., & Tang, W. (2020). Pricing research of long-term care insurance based on non-homogeneous Markov model. Insurance Studies, (07), 108-121. DOI:10.13497/j.cnki.is.2020.07.009.

[2] Qiu, C. J., Liu, S. X., & Zhang, N. (2023). Research on end-to-end long-term care insurance pricing model based on deep neural network. Insurance Studies, (12), 71-81. DOI:10.13497/j.cnki.is.2023.12.006.

[3] Wang, X. J., & Wang, J. Y. (2018). Pricing of long-term care insurance based on Markov model. Insurance Studies, (10), 87-99. DOI:10.13497/j.cnki.is.2018.10.008.

[4] Qiu, C. J., Guan, H. L., Qian, L. Y., et al. (2020). Pricing research of long-term care insurance—Based on XGBoost algorithm and BP combined neural network model. Insurance Studies, (12), 38-53. DOI:10.13497/j.cnki.is.2020.12.003.

[5] Cheng, G. P., Shen, S. J., & Xu, D. N. (2024). Pricing strategy of China's long-term care insurance products based on combined machine learning model. Insurance Studies, (12), 57-71. DOI:10.13497/j.cnki.is.2024.12.005.

[6] Duan, Y. X., Peng, X. M., & Fang, K. N. (2024). Financial early warning research of Chinese property insurance companies—Based on random forest financial diagnosis method. Insurance Studies, (04), 20-33. DOI:10.13497/j.cnki.is.2024.04.002.