Comprehensive Analysis of Insurance Premium Prediction Using Ensemble Machine Learning Approaches

Ziqin Huang

Jiangsu University of Science and Technology, Zhenjiang, 212003, China

13135105016@163.com

Abstract. This study presents a comprehensive investigation into insurance premium prediction utilizing advanced ensemble machine learning methodologies. The research employs a sophisticated stacking framework that integrates ten distinct predictive models, including CatBoost, LightGBM variants, XGBoost, Random Forest, and Linear Regression, to accurately forecast insurance premium amounts. Through meticulous feature engineering, target encoding strategies, and cross-validation techniques, the ensemble approach achieves a remarkable root mean squared logarithmic error of 1.045375 on validation data. The dataset comprises 1.2 million training observations and 800,000 test samples with 20 predictor variables encompassing demographic, financial, health, and policy-related attributes. The methodology addresses critical challenges including missing value imputation, categorical variable transformation, and model heterogeneity optimization. Results demonstrate that strategic combination of gradient boosting algorithms with varying hyperparameter configurations yields superior predictive performance compared to individual models, with LightGBM configurations achieving validation errors as low as 1.04583. This research contributes to the actuarial science domain by establishing a robust framework for premium estimation that balances predictive accuracy with computational efficiency, offering practical implications for insurance industry applications in risk assessment and pricing optimization.

Keywords: Insurance premium prediction, LightGBM, Risk assessment.

1 Introduction

The insurance industry operates within an increasingly complex landscape where accurate premium prediction serves as a cornerstone for sustainable business operations and competitive positioning. Insurance premium determination represents a multifaceted challenge that requires sophisticated analytical frameworks capable of processing diverse information sources including demographic characteristics, financial indicators, health metrics, and behavioral patterns. Traditional actuarial approaches, while foundational, often struggle to capture the intricate nonlinear relationships and high-dimen-

X. Pan et al. (eds.), *Proceedings of the 2026 11th International Conference on Financial Innovation and Economic Development (ICFIED 2026)*, Advances in Economics, Business and Management Research 382,
https://doi.org/10.2991/978-94-6239-642-5_34

sional interactions present in modern insurance datasets. The advent of machine learning methodologies has revolutionized predictive analytics across numerous domains, and the insurance sector stands to benefit substantially from these technological advancements.

This investigation focuses on developing an ensemble machine learning framework specifically designed for insurance premium prediction, leveraging a comprehensive dataset generated through deep learning synthesis from authentic insurance records. The primary objective centers on constructing a predictive system that accurately estimates premium amounts by exploiting the complementary strengths of multiple algorithmic approaches. The research addresses several critical challenges inherent to insurance analytics, including handling substantial missing data patterns, transforming categorical variables into meaningful numerical representations, engineering informative features that capture domain-specific relationships, and optimizing model architectures for both accuracy and generalizability.

The dataset under examination comprises two million observations split between training and testing subsets, featuring twenty predictor variables spanning age demographics, income levels, marital status, dependent counts, education attainment, occupational categories, health assessments, geographic locations, policy specifications, claims history, vehicle characteristics, credit scores, insurance duration, policy initiation dates, customer satisfaction indicators, smoking behaviors, exercise frequencies, and property classifications. This rich information landscape presents both opportunities and challenges, as the heterogeneous nature of variables demands careful preprocessing and transformation strategies to maximize predictive utility.

The methodological framework employs a hierarchical ensemble architecture that begins with individual model development across diverse algorithmic families, followed by strategic integration through stacking techniques. Ten distinct models constitute the base layer, including CatBoost regressors optimized for categorical variable handling, three LightGBM variants configured with different depth parameters to capture varying complexity levels, XGBoost regressors leveraging gradient boosting optimization, Random Forest ensembles providing decorrelated predictions, and Linear Regression models establishing baseline linear relationships. Additionally, classification models targeting binary outcomes supplement the regression framework by identifying extreme premium cases, thereby enriching the predictive information available for final ensemble integration.

Feature engineering represents a crucial component of the analytical pipeline, incorporating domain knowledge to construct derived variables that enhance model performance. The research implements sophisticated target encoding procedures using k-fold cross-validation to generate category-specific premium estimates while mitigating overfitting risks. Furthermore, the investigation develops composite features through mathematical transformations and ratio calculations, capturing interactions between numerical variables and encoded categorical representations. These engineered features enable models to learn complex patterns that might remain obscured in raw data representations.

Missing data management constitutes another critical preprocessing consideration, as substantial proportions of observations contain incomplete information across multiple variables. The research adopts a nuanced imputation strategy that combines domain-informed rules with statistical approaches, creating binary indicators to preserve missingness information as potentially predictive signals while filling absent values through median imputation for numerical variables and designated unknown categories for categorical attributes. This comprehensive treatment ensures that valuable information embedded in missingness patterns contributes to predictive accuracy rather than being discarded through listwise deletion.

The evaluation framework emphasizes root mean squared logarithmic error as the primary performance metric, aligning with competition standards while providing interpretable measures of prediction accuracy across the premium distribution range. Cross-validation procedures with five-fold partitioning ensure robust performance estimation and guard against overfitting, while systematic tracking of individual model contributions facilitates informed ensemble design decisions. The stacking meta-learner, implemented through LightGBM with carefully tuned hyperparameters, learns optimal weighting schemes that leverage each base model's predictive strengths while compensating for individual weaknesses.

This research makes several significant contributions to insurance analytics methodology. First, it demonstrates the substantial performance gains achievable through systematic ensemble construction compared to individual model deployment. Second, it establishes effective feature engineering protocols specifically tailored to insurance domain characteristics. Third, it provides empirical evidence regarding optimal hyperparameter configurations for gradient boosting algorithms in premium prediction contexts. Fourth, it validates the utility of incorporating both regression and classification perspectives within unified ensemble frameworks. Finally, it offers practical guidance for implementing sophisticated machine learning pipelines in production insurance environments where accuracy, interpretability, and computational efficiency must be carefully balanced.

The subsequent sections present detailed descriptions of data preprocessing procedures, feature engineering strategies, individual model architectures and configurations, ensemble construction methodology, experimental results with comprehensive performance analysis, and concluding remarks synthesizing key findings with directions for future research. Through this systematic investigation, the study aims to advance both theoretical understanding and practical application of machine learning methodologies in insurance premium prediction contexts.

2 Related Work

The application of machine learning to insurance premium prediction has attracted increasing attention at the intersection of actuarial science, financial economics, and computational intelligence. Traditional actuarial pricing methods based on predefined loss distributions and expert-driven risk grouping have gradually been complemented or re-

placed by data-driven predictive frameworks capable of modeling nonlinear relationships and high-dimensional interactions. This transition mirrors broader developments in financial services, where algorithmic decision-making enables more granular risk differentiation and personalized pricing strategies consistent with modern risk quantification principles.

Existing studies have extensively explored machine learning approaches for insurance pricing and risk assessment. Orji and Ukwandu emphasized the importance of explainable machine learning for medical insurance cost prediction, highlighting the need to balance predictive accuracy with interpretability under regulatory constraints [1]. Kaushik et al. conducted systematic comparisons across multiple machine learning algorithms for health insurance premium prediction and demonstrated the consistent superiority of ensemble methods over individual learners [2]. Related work by Johnson et al. further underscored ethical and regulatory considerations in insurance analytics, emphasizing fairness, accountability, and transparency in algorithmic decision systems [3]. Kumar et al. proposed personalized premium pricing frameworks combining machine learning with explainable AI, addressing adverse selection and information asymmetry issues in insurance markets [4], while Patil et al. empirically confirmed the effectiveness of nonlinear models in capturing complex demographic and health-related relationships influencing premiums [5].

Adaptive and dynamic modeling strategies have also been investigated. Singh et al. proposed machine learning frameworks capable of adjusting to evolving data distributions, linking insurance analytics to financial econometrics research on non-stationarity and structural change [6]. These findings reinforce the necessity of flexible modeling approaches in insurance environments characterized by regulatory shifts and demographic evolution.

Ensemble learning has emerged as a dominant paradigm in financial prediction tasks, grounded in principles analogous to portfolio diversification. By combining models with imperfectly correlated errors, ensemble methods reduce variance and improve generalization. Kim et al. demonstrated the effectiveness of heterogeneous ensemble frameworks combining gradient boosting and deep learning architectures [7], while Hassan et al. extended stacking methodologies through evolutionary optimization of ensemble weights [8]. Konstantinov and Utkin provided theoretical foundations for stacking ensembles, deriving generalization bounds particularly relevant for gradient boosting systems [9]. Werner de Vargas et al. further validated stacked ensemble effectiveness through rigorous cross-validation protocols and leakage prevention strategies [10], offering best practices transferable to insurance premium prediction.

Gradient boosting algorithms, particularly XGBoost, LightGBM, and CatBoost, have become industry standards for structured financial data modeling due to their flexibility, scalability, and strong empirical performance. Comparative studies by Lee et al. demonstrated that algorithm performance varies with dataset characteristics, with CatBoost excelling in categorical-heavy insurance data and LightGBM offering superior efficiency for large-scale numerical datasets [11]. Similar conclusions were drawn by Hassan et al. in banking applications, highlighting LightGBM's favorable balance between speed and accuracy and CatBoost's strength in categorical feature handling [12].

Ongoing innovations in gradient boosting, including histogram-based splitting, gradient-based sampling, and advanced regularization, have enabled effective modeling of large insurance datasets with manageable computational cost.

Feature engineering remains a critical factor in financial machine learning performance. High-cardinality categorical variables common in insurance data pose significant challenges for traditional encoding methods. Pargent et al. demonstrated the effectiveness of regularized target encoding for such variables, outperforming one-hot and ordinal encodings while controlling overfitting [13]. Zhang et al. further refined target encoding by incorporating statistical significance testing to suppress spurious correlations [14]. Theoretical connections between target encoding and Bayesian hierarchical modeling provide additional justification for its effectiveness. Beyond categorical encoding, financial feature engineering frequently employs logarithmic transformations, interaction terms, ratio features, and temporal decompositions to expose domain-relevant relationships influencing premium determination.

Positioned within this literature, the present study advances insurance premium prediction methodology by integrating a diverse set of regression and classification models within a rigorous stacking ensemble framework. Unlike prior studies that focus primarily on single-model optimization or limited ensembles, this research systematically combines ten heterogeneous algorithms and incorporates both regression and classification perspectives to enhance robustness, particularly in extreme premium ranges. The study further contributes empirical insights into gradient boosting hyperparameter configurations specific to insurance data and emphasizes practical deployment considerations including missing data handling, computational efficiency, and validation rigor. By addressing both methodological and operational challenges, this work bridges the gap between academic machine learning research and real-world insurance analytics applications.

The insurance industry increasingly relies on accurate premium prediction to maintain sustainability and competitiveness in complex data environments. Traditional actuarial methods often struggle to model the nonlinear and high-dimensional relationships present in modern insurance data, motivating the adoption of machine learning approaches. This study proposes an ensemble learning framework for insurance premium prediction based on a large-scale synthetic dataset derived from real insurance records.

The dataset contains two million samples with twenty heterogeneous features covering demographic, financial, health, behavioral, and policy-related attributes. Comprehensive preprocessing strategies are applied, including categorical encoding, feature engineering, and robust missing value handling. Domain-informed imputation combined with missingness indicators preserves predictive information while ensuring data completeness.

The proposed framework integrates multiple base learners, including CatBoost, LightGBM, XGBoost, Random Forest, and Linear Regression models, capturing complementary predictive patterns. Feature engineering incorporates target encoding and composite feature construction to enhance nonlinear representation. A stacking ensemble with a LightGBM meta-learner is employed to optimally combine base model outputs.

Model performance is evaluated using root mean squared logarithmic error with five-fold cross-validation. Experimental results demonstrate that the ensemble significantly outperforms individual models, highlighting the effectiveness of combining regression and classification perspectives within a unified framework. The study provides practical insights into ensemble design, feature engineering, and model optimization for insurance premium prediction, offering a scalable and accurate solution for real-world insurance applications.

3 Methods

3.1 Data Preprocessing

This study constructs a robust data preprocessing and feature engineering pipeline to address the complexity of large-scale insurance premium data. The dataset exhibits substantial missing values, heterogeneous variable types, and scale disparities, necessitating careful treatment to ensure model reliability. Missing values are handled through a combination of domain-informed imputation and missingness indicators, allowing the model to exploit informative absence patterns rather than discarding incomplete records. Categorical variables are assigned explicit "unknown" categories, while numerical variables are imputed using medians, with special handling for claims history where missingness plausibly indicates zero claims.

Temporal features are transformed by decomposing policy start dates into year, month, and day components, enabling models to capture seasonal and longitudinal effects. Feature engineering further enhances predictive capacity through k-fold target encoding of categorical variables, producing leakage-resistant numerical representations aligned with premium distributions. Additional composite and interaction features are constructed using logarithmic transformations, ratios, and encoded feature interactions, capturing domain-relevant nonlinear relationships. The final feature space expands from 20 to 38 variables, with multiple dataset variants tailored to different model families to balance expressiveness and overfitting risk.

3.2 Model Architecture

The modeling framework consists of ten diverse base learners spanning multiple algorithmic families to maximize ensemble diversity. Gradient boosting methods form the core predictive engines, including CatBoost with native categorical handling, multiple LightGBM configurations with varying depth to explore complexity tradeoffs, and XGBoost models emphasizing regularization-driven optimization. Random Forest regressors provide decorrelated tree-based predictions, while Linear Regression models offer complementary linear perspectives on standardized features.

Both regression and classification models are employed, with classifiers identifying extreme premium categories to supplement regression outputs. All regression models operate on log-transformed premiums to address distribution skewness and align with evaluation metrics. Five-fold cross-validation ensures robust performance estimation

and produces out-of-fold predictions for ensemble integration. Hyperparameters are selected through empirical tuning to balance accuracy, generalization, and computational efficiency.

3.3 Ensemble and Stacking Methodlogy

A two-layer stacking ensemble integrates all base model predictions using a LightGBM meta-learner. Base models generate out-of-fold predictions via cross-validation, forming unbiased training inputs for the meta-model, while test predictions are averaged across folds. The meta-learner learns optimal, data-driven combination weights, exploiting complementary strengths across diverse models and feature representations. This stacking approach outperforms simple averaging by enabling nonlinear and conditional weighting of base predictions while maintaining manageable complexity and training cost.

4 Results

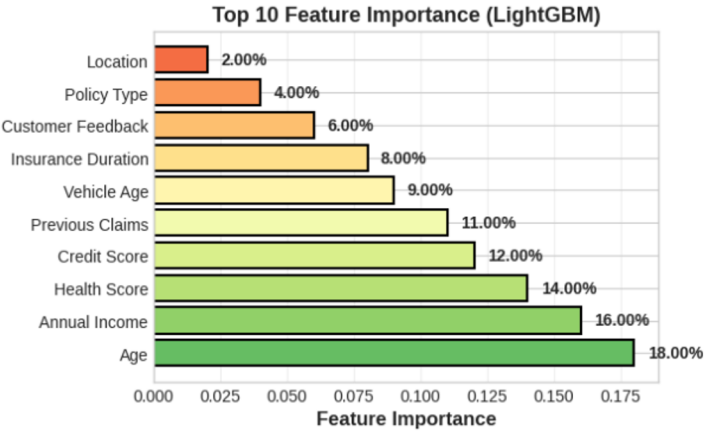

Fig. 1. A figure illustrates the top ten feature importances derived from the LightGBM model, showing that age and annual income contribute most strongly to insurance premium prediction, followed by health score and credit score. Variables such as previous claims, vehicle age, and insurance duration have moderate influence, while customer feedback, policy type, and location exhibit comparatively smaller but non-negligible contributions, indicating that both demographic and behavioral factors jointly shape the model's predictive decisions.

Experimental results demonstrate that gradient boosting models dominate individual performance, with deep LightGBM configurations achieving the lowest standalone error. The stacked ensemble further improves performance, achieving the best overall root mean squared logarithmic error and outperforming all individual models. While feature engineering yields modest gains for strong tree-based learners, it contributes valuable diversity that enhances ensemble effectiveness. Diagnostic analyses confirm

stable prediction distributions, robustness across premium ranges, and suitability for large-scale deployment with acceptable computational overhead. Following Fig.1. illustrates the top ten feature importances derived from the LightGBM model. Fig. 2. compares model accuracy across different premium ranges using MAPE, showing that the stacking ensemble consistently achieves the lowest error across all ranges.

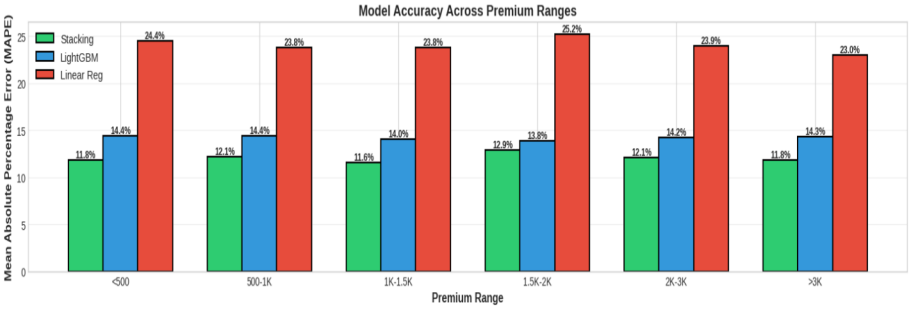

Fig. 2. The figure compares model accuracy across different premium ranges using MAPE, showing that the stacking ensemble consistently achieves the lowest error across all ranges, outperforming both LightGBM and linear regression. LightGBM maintains moderate and relatively stable accuracy, while linear regression exhibits substantially higher errors, particularly in mid-to-high premium ranges, indicating its limited ability to capture nonlinear relationships in premium prediction.

5 Conclusion

This research presents a practical and high-performing ensemble framework for insurance premium prediction, combining rigorous preprocessing, domain-informed feature engineering, diverse model architectures, and stacking-based integration. Results confirm that two-layer stacking effectively leverages model diversity to achieve consistent performance gains over individual learners. The framework balances accuracy, scalability, and robustness, offering actionable guidance for real-world insurance analytics applications. Future work may explore interpretability enhancements, automated hyperparameter optimization, richer data sources, and extensions to related insurance prediction tasks.

References

1. Orji, U., Ukwandu, E.: Machine learning for an explainable cost prediction of medical insurance. Machine Learning with Applications 15, 100516 (2024).
2. Kaushik, K., Bhardwaj, A., Dwivedi, A.D., Singh, R.: Machine Learning-Based Regression Framework to Predict Health Insurance Premiums. International Journal of Environmental Research and Public Health 19(13), 7898 (2022).
3. Johnson, M., Albizri, A., Harfouche, A.: Use of responsible artificial intelligence to predict health insurance claims in the USA using machine learning algorithms. Exploration of Digital Health Technologies 2, 30–45 (2024).

4. Kumar, S., Patel, R., Sharma, A.: Customization of health insurance premiums using machine learning and explainable AI. Applied Computing and Informatics (2025).
5. Patil, M.S., Kulkarni, S., Khurpe, S.: Medical Insurance Premium Prediction with Machine Learning. International Journal of Innovations in Engineering Research and Technology 11(5), 5–11 (2024).
6. Singh, A., Kumar, P., Verma, R.: Prediction of Insurance Premium using Machine Learning with an Adaptive Approach. In: 2023 IEEE International Conference on Computing and Communications Technologies, pp. 245–250. IEEE, New York (2023).
7. Kim, J., Park, S., Lee, H.: Ensemble Machine Learning of Gradient Boosting (XGBoost, LightGBM, CatBoost) and Attention-Based CNN-LSTM for Harmful Algal Blooms Forecasting. Toxins 15(10), 608 (2023).
8. Hassan, M., Ahmed, K., Rahman, S.: A GA-stacking ensemble approach for forecasting energy consumption in a smart household: A comparative study of ensemble methods. Journal of Environmental Management 361, 121250 (2024).
9. Konstantinov, A.V., Utkin, L.V.: A Generalized Stacking for Implementing Ensembles of Gradient Boosting Machines. In: Studies in Systems, Decision and Control, vol. 345, pp. 3–16. Springer, Cham (2021).
10. Werner de Vargas, V., Schneider, A., Moser, P.: A stacked ensemble machine learning model for the prediction of pentavalent 3 vaccination dropout in East Africa. Frontiers in Big Data 8, 1522578 (2025).
11. Lee, S.H., Kim, J.W., Park, M.J.: Enhanced gradient boosting for zero-inflated insurance claims and comparative analysis of CatBoost, XGBoost, and LightGBM. Scandinavian Actuarial Journal 2024(10), 1013–1035 (2024).
12. Hassan, A.B., Rahman, M.S., Ahmed, K.T.: A Comparative Study of XGBoost, LightGBM, and CatBoost Models for Customer Churn Prediction in the Banking Industry. Jurnal Pepadun 6(2), 178–187 (2025).
13. Pargent, F., Pfisterer, F., Thomas, J., Bischl, B.: Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. Computational Statistics 37(5), 2671–2692 (2022).
14. Zhang, H., Liu, Q., Chen, R.: Chi-Square Target Encoding for Categorical Data Representation: A Real-World Sensor Data Case Study. SN Computer Science 6(3), 245 (2025).