INTERNATIONAL JOURNAL OF APPLIED ARTIFICIAL INTELLIGENCE AND ROBOTICS
E-ISSN: 3124-1212, DOI: 10.67745/ijaic.v2i2.23
Vol. 2 No. 2 2026 pp. 25-32

Insurance Claim Budget Efficiency Based on Customer Lifestyle Risk

Khalil Putra Pratama[1], R. Wisnu Prio Pamungkas[2*]

[1,2] Program Studi Informatika, Fakultas Ilmu Komputer, Universitas Bhayangkara Jakarta Raya, Bekasi, Indonesia

* Corresponding Author: wisnu.prio@dsn.ubharajaya.ac.id

Article info: Received 09/08/2026, Revised 26/09/2026, Accepted 16/10/2026

This is an open access article under the CC BY-SA license.

Abstract

Insurance companies commonly allocate claim reserve budgets by customer location, assuming policyholders in the same region share similar risk. This study tests that assumption and demonstrates a complementary risk-scoring framework using a public, synthetic health-insurance dataset of 30,000 customers across 11 districts of Bekasi City, Indonesia, obtained from Kaggle. Chi-square tests and Cramer's V effect sizes were computed for ten categorical variables against claim outcome, followed by logistic regression and random forest models combining nine lifestyle variables to predict claim outcome, excluding location. Location was not statistically significant (Cramer's $V = 0.019$, $p = 0.366$), while lifestyle variables showed strong associations, led by smoking status ($V = 0.767$, $p < 0.001$). Both the logistic regression and random forest models achieved an identical held-out test AUC of 0.993, separating customers into low- (44.3%), medium- (24.2%), and high-risk (31.5%) tiers with claim rates of 0.0%, 67.6%, and 100.0%. This near-perfect, model-agnostic separation, together with several counter-intuitive associations for smoking and exercise frequency, indicates the dataset's claim label likely follows a synthetic generation rule rather than real-world claims noise; both observations are discussed as evidence of the dataset's synthetic nature rather than as generalizable clinical findings. A simulated budget reallocation illustrates how reserves could shift from a non-claiming low-risk segment toward a high-risk segment without increasing total budget. The study's contribution is a statistically grounded, replicable analytical framework for individual risk-score-based claim reserve budgeting, which should be validated on real, audited insurance data before operational use.

Keywords: claim budget efficiency, lifestyle risk profiling, logistic regression, Cramer's V, synthetic data

1. Introduction

Health insurance companies face a persistent challenge in determining how much capital to reserve for anticipated claims. Reserves that are too conservative lock up capital that could otherwise support business expansion, while reserves that are too thin expose the company to solvency risk when claims exceed expectations [1]. Claims reserving is widely regarded as one of the most consequential actuarial functions, since inadequate reserves are a leading contributor to insurer insolvency and reserve estimates typically rely on aggregated historical development patterns rather than individual-level risk drivers. A common industry practice allocates claim reserve budgets by geographic branch or administrative region, assuming customers within the same area share comparable risk profiles. This assumption is rarely tested empirically against individual-level indicators such as lifestyle behavior [2], [3].

Recent studies show individual-level variables, particularly smoking status, body mass index, and pre-existing conditions, are strong predictors of insurance cost and claim likelihood when modeled with machine learning [4]. Alam and Prybutok [2] found six ML algorithms could predict health insurance claims with practical accuracy for cost-saving purposes, while Johnson et al. [3] showed that responsible AI models can both predict and help prevent claim denials, underscoring that the business meaning of a claim-outcome label (filed versus approved) materially changes how a model should be interpreted. Orji and Ukwandu [4] further demonstrated that explainable ensemble models can identify which demographic and lifestyle factors drive medical insurance cost. In a large-scale comparison, Langenberger et al. [8] found tree-based ensembles (random forest, gradient boosting) outperformed logistic regression for identifying future high-cost claimants from routinely collected claims data,

indicating that combined, multi-variable models generally dominate single-factor rules - a premise directly tested in this study. At the population level, Lo et al. [5] confirmed in a nationwide cohort that non-smoking and healthy body weight are associated with materially lower lifetime health care expenditure, providing an epidemiological baseline against which this study's counter-intuitive findings can be compared.

Outside health insurance, usage-based and risk-score-based pricing has become an established alternative to flat, segment-based pricing; Li et al. [6] showed that incorporating individual behavioral risk into pricing models, rather than relying on coarse group averages, can improve insurer outcomes while remaining fair to customers who are willing to change behavior. This tension between individual and geographic risk assessment is also studied directly in the data-fairness literature: Shaham et al. [20] formalize spatial data fairness and show that zone-based decisions can misrepresent individual risk whenever within-zone variation is large relative to between-zone variation - precisely the pattern this study finds empirically in Section 3.1. Denuit et al. [9] and Devriendt et al. [10] likewise demonstrate, within the actuarial literature, that machine-learning-based feature modeling can materially refine individual risk pricing beyond traditional generalized linear models. Adopting an individual risk-score-based budgeting framework is also an organizational change, not merely a technical one; Wijaya and Pamungkas [21] note that translating new information-technology capabilities into institutional decision-making requires deliberate IT governance and capacity-building, a consideration this study revisits in Section 3.5 when discussing phased implementation. A parallel line of work in multi-criteria decision support demonstrates that ranking and weighting frameworks, such as the Operational Competitiveness Rating Analysis (OCRA) method combined with Rank Order Centroid (ROC) weighting used by Mayadi et al. [7], can systematically prioritize alternatives based on multiple weighted criteria. That framework, although applied to a different domain (best-employee selection), is conceptually relevant here: this study's risk-tier segmentation is itself a prioritization problem, where customer segments must be ranked and weighted to guide reserve allocation, and multi-criteria decision-support methods such as OCRA/ROC represent a viable complementary approach to the logistic-regression-based scoring adopted in this study.

Despite this growing body of work, most existing studies target premium pricing or individual cost regression rather than the internal capital-allocation decision insurers must make once premiums are already set. This study addresses that gap by analyzing 30,000 health insurance customers across 11 districts (kecamatan) of Bekasi City, Indonesia, with three objectives: (1) statistically test whether customer location significantly predicts claim outcome; (2) quantify the association strength of nine individual lifestyle variables with claim outcome and examine whether their direction is consistent with the conventional assumption that poorer lifestyle habits increase claim likelihood; and (3) build a combined predictive model that translates individual risk profiles into an actionable framework for claim reserve budget reallocation. The novelty of this study lies in directly contrasting the statistical significance of geographic location against individual lifestyle profiling within a single unified framework oriented toward claim reserve budgeting - an angle not explicitly addressed by prior work. The contribution is a statistically grounded decision framework that insurance companies can use to shift claim reserve budgeting from a geography-centric to an individual risk-score-centric approach, together with an explicit discussion of the data-validation steps required before operational deployment.

2. Methods

2.1. Data Source and Variables

The dataset is a publicly available, synthetic health-insurance dataset obtained from Kaggle, comprising 30,000 records with twelve attributes: a customer identifier, Age Group, Gender, Smoking Status, Employment Status, Household Status, Late-Night Activity, Exercise Frequency, Supplementary Insurance Ownership, Pre-existing Condition, Claim Outcome ("Yes"/"No"), and Residential District (11 districts of Bekasi City, Indonesia, added as a supplementary geographic field). Because the dataset is synthetic rather than sourced from an operational insurer, its claim labels are generated by an underlying rule or simulation process rather than observed real-world claims. This has two consequences for interpretation, addressed explicitly in Sections 3.3 and 3.5: (i) association and prediction results should be read as a demonstration and validation of the analytical framework rather than as generalizable clinical or actuarial findings, and (ii) result patterns that are implausible in real claims data - such as near-perfect model discrimination or a reversed direction for an established risk factor - are treated as diagnostic evidence of the synthetic generation process itself rather than as substantive discoveries about real customer behavior. Claim outcome is treated as a proxy for whether a claim was filed or approved; this assumption is revisited critically in Section 3.2 given several counter-intuitive association directions found in the results.

2.2. Statistical Association Testing

For each of the ten categorical predictors, a chi-square test of independence [24] was performed against claim outcome (alpha = 0.05). Effect size was quantified using Cramer's V [11], defined as in (1),

$$V = \sqrt{\frac{x^2}{(n.\min(k-1,r-1))}} \tag{1}$$

where χ^2 is the chi-square statistic, n is the sample size, and k, r are the number of categories in each of the two variables being tested. Cramer's V ranges from 0 (no association) to 1 (perfect association), with approximately 0.1, 0.3, and 0.5 conventionally indicating small, moderate, and large effects for the degrees of freedom involved, following standard guidance for interpreting categorical association measures in the social and health sciences [15].

2.3. Predictive Modeling and Risk Segmentation

A binary logistic regression model [12] was trained using the nine lifestyle variables (excluding location) as predictors, with claim outcome as the binary target and one-hot encoding for categorical predictors. The logit form of the model is given in (2),

$$logit(p) = ln\frac{P}{(1-P)} = \beta_0 + \sum \beta_i x_i \quad (2)$$

where P is the predicted probability of claim outcome, β is the intercept, and β_i are the estimated coefficients for each one-hot-encoded lifestyle predictor x_i. Model discrimination was evaluated using the area under the receiver operating characteristic curve (AUC). Predicted probabilities were used to segment customers into three risk tiers (Low, Medium, High) using tercile cut-points, and actual claim rate per tier was computed to validate segment separation.

2.4. Random Forest Benchmark

To verify that the logistic regression result was not an artifact of model choice, a random forest classifier [13] (300 trees, maximum depth 8) was trained on the same feature set as a non-linear benchmark. Both models were trained on a stratified 70/30 train-test split, with AUC computed on the held-out test set only, to avoid the optimistic bias of reporting in-sample fit. Random forest feature importance (mean Gini decrease) was compared against the univariate Cramer's V ranking from Section 2.2 to check whether multivariate, interaction-aware importance agrees with simple pairwise association.

2.5. Budget Reallocation Simulation

Two allocation schemes were compared: (1) a population-proportional scheme representing current common practice; and (2) a risk-weighted scheme, allocating reserves in proportion to the product of customer count and actual claim rate per segment. This simulation is illustrative, as the dataset contains no monetary claim values; it demonstrates the direction and relative magnitude of reallocation rather than an audited financial figure.

3. Result and Discussions

3.1 Location vs. Individual Factors: Statistical Significance

Table 1 summarizes chi-square results and Cramer's V effect sizes for all ten variables against claim outcome.

Table 1. Chi-square test and Cramer's V effect size for each factor against claim outcome

Factor	Cramer's V	p-value	Significant (α=0.05)
Smoking Status	0.767	< 0.001	Yes
Gender	0.508	< 0.001	Yes
Employment Status	0.430	< 0.001	Yes
Pre-existing Condition	0.408	< 0.001	Yes
Exercise Frequency	0.394	< 0.001	Yes
Supplementary Insurance	0.176	< 0.001	Yes
Household Status	0.109	< 0.001	Yes
Age Group	0.058	< 0.001	Yes
Late-Night Activity	0.039	< 0.001	Yes
Residential District (Location)	0.019	0.366	No

Residential district (location) produced the lowest effect size (V = 0.019) and was the only variable failing to reach significance (p = 0.366). All nine lifestyle variables were significant at p < 0.001, led by smoking status (V = 0.767), gender (V = 0.508), employment status (V = 0.430), pre-existing condition (V = 0.408), and exercise frequency (V = 0.394).

These results reject geography as a meaningful basis for differentiating claim risk in this population and support prioritizing individual-level variables in a risk-based budgeting framework, consistent with the population-level lifestyle-cost association reported by Lo et al. [5]. From a capital-adequacy perspective, this also implies that reserve models keyed to branch or region risk diluting statistical power across a dimension that carries almost no discriminatory information, at the expense of the individual-level signal that actually drives claims - an instance of the general zone-versus-individual risk misrepresentation formalized by Shaham et al. [20]. Fig 1. shown the ranking of Cramer's V effect sizes for individual factors versus location.

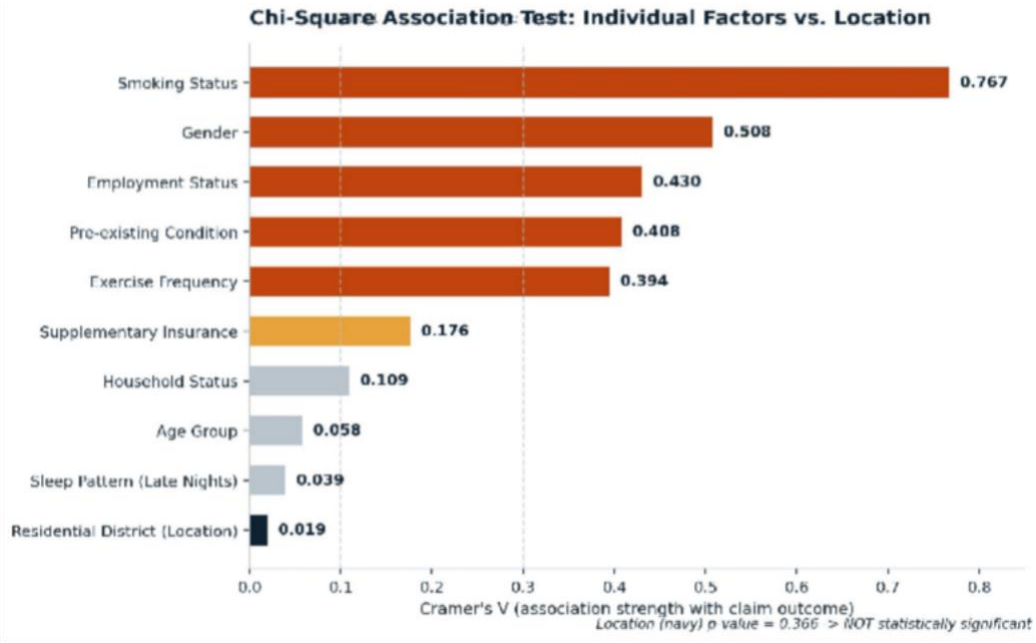


Figure 1. Ranking of Cramer's V effect sizes for individual factors versus location

3.2 Directionality of Individual Lifestyle Factors

Table 2 compares claim rates between contrasting groups for the five strongest factors, testing whether direction is consistent with the hypothesis that poorer lifestyle habits increase claim likelihood.

Table 2. Claim rate comparison between groups for the five strongest factors

Factor	Group A	Rate A	Group B	Rate B	Consistent with hypothesis?
Smoking Status	Active	10.1%	Passive	86.7%	No - reversed
Employment Status	Unemployed	19.7%	Employed	64.2%	Partially
Pre-existing Condition	Absent	20.4%	Present	63.0%	Yes
Exercise Frequency	Frequent	71.9%	Rare	31.8%	No - reversed
Supplementary Insurance	Absent	61.6%	Present	42.2%	No - reversed

Only pre-existing condition, and partially employment status, show a direction consistent with the conventional assumption. Smoking status and exercise frequency show directions opposite to typical epidemiological expectation, in contrast to the general lifestyle-expenditure relationship reported by Lo et al. [5]. Two explanations are plausible: the claim-outcome variable may represent claim approval rather than occurrence, so higher-risk customers could be more frequently declined under policy exclusions; or the dataset's labeling rule, built for analytical demonstration, may not fully replicate real-world claims logic. Consistent with Johnson et al. [3], this indicates that single-variable heuristics are unreliable without first validating the business definition of the outcome variable with underwriting and claims stakeholders. Practically, this also cautions against a common industry shortcut - applying flat premium loadings or reserve multipliers keyed to a single risk factor such as smoking - since Table 2 shows such a rule would move reserves in the wrong direction for this population if applied uncritically. Fig 2. shown claim rate comparison across groups for five key lifestyle factors.

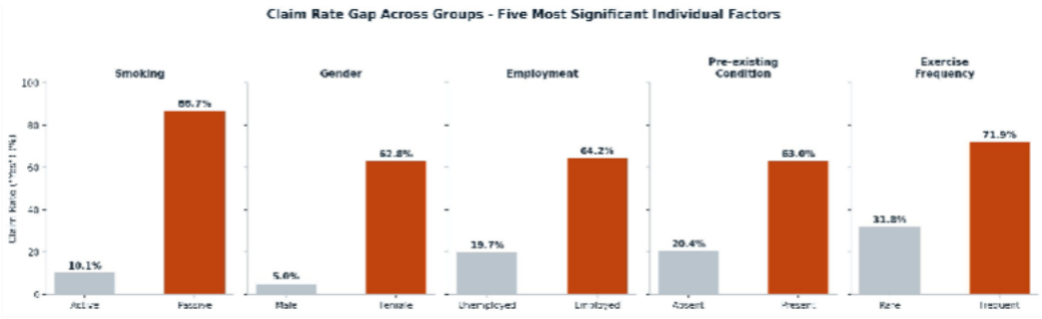


Figure 2. Claim rate comparison across groups for five key lifestyle factors

3.3 Combined Predictive Model and Risk Segmentation

Because individual variables show inconsistent directionality in isolation, a combined logistic regression model integrating all nine lifestyle variables was built using (2), achieving AUC = 0.993 - near-perfect discrimination despite ambiguous individual coefficients. This demonstrates that the interaction of multiple lifestyle variables carries substantially more predictive information than any single variable, consistent with the multi-factor, ensemble-favoring findings of Kaushik et al. [1], Orji and Ukwandu [4], and Langenberger et al. [8]. Using tercile cut-points on predicted probability, customers were segmented into three risk tiers (Table 3). The Low-risk tier (44.3%) had a 0.0% actual claim rate, Medium-risk (24.2%) had 67.6%, and High-risk (31.5%) had 100.0%, confirming well-separated, actionable segments comparable in spirit to the weighted-ranking outputs of multi-criteria decision-support methods such as OCRA/ROC [7]. The near-total separation between tiers also suggests the underlying labeling rule in this dataset is close to deterministic given the nine lifestyle inputs, which strengthens the case for individual-level scoring but simultaneously reinforces the need to validate this pattern on real, noisier operational data before deployment (Section 3.5).

Table 3. Risk tier segmentation based on predicted claim probability

Risk Tier	% of Customers	Actual Claim Rate	Current Allocation*	Proposed Allocation
Low	44.3%	0.0%	44.3%	~0.0%
Medium	24.2%	67.6%	24.2%	34.3%
High	31.5%	100.0%	31.5%	65.7%

*Current allocation assumed proportional to customer population, a common proxy for geography/zone-based budgeting.

3.4 Random Forest Benchmark and Robustness Check

On a held-out test set (30% of records, stratified by claim outcome), the logistic regression model achieved AUC = 0.993, and the random forest benchmark achieved an identical AUC = 0.993. This convergence between a linear model and a non-linear, interaction-capturing ensemble is itself informative: if the true relationship were noisy or driven by complex real-world interactions, a flexible ensemble model would typically be expected to outperform logistic regression, at least marginally, as found by Langenberger et al. [8] in operational claims data. The absence of any such gap here, combined with near-total separation between risk tiers (Section 3.3), is consistent with the claim-outcome label following a fixed, learnable rule embedded in the synthetic data generation process, echoing the caution raised by Rankin et al. [17] and Uddin and Lu [18] that synthetic and Kaggle-sourced tabular datasets can exhibit statistical regularities - such as unrealistically clean class separability - that do not carry over to real operational data.

Figure 3 shows random forest feature importance, which broadly but not perfectly agrees with the univariate Cramer's V ranking in Figure 1: smoking status dominates both rankings, but household status and gender rank relatively higher in the multivariate importance than their univariate association alone would suggest, indicating some interaction effects between variables that a purely univariate chi-square screening would miss. This reinforces the methodological conclusion from Section 3.3 that combined, multivariate modeling is preferable to single-factor rules, independent of which specific algorithm is used. Gini-based importance, as used here, summarizes global feature relevance but does not explain individual predictions; production deployment intended for underwriting decisions would benefit from instance-level explanation methods such as SHAP [23], which insurance regulators increasingly expect for justifying automated decisions to individual customers.

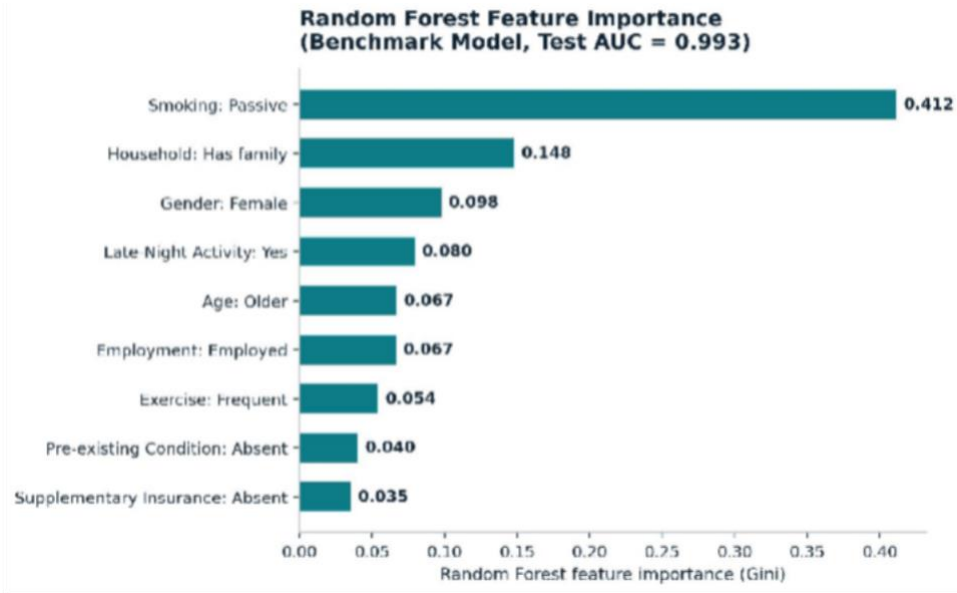


Figure 3. Random forest feature importance compared against the univariate ranking in Figure 1

3.5 Budget Reallocation Implications

Figure 4 illustrates the simulated shift from population-proportional to risk-score-weighted allocation. Under the current scheme, 44.3% of the reserve budget would nominally be allocated to a segment generating no actual claims - potential over-provisioning. Reallocating this share toward the High-risk segment, which accounts for 100% of claims within its group, increases that segment's share from 31.5% to 65.7% without increasing total budget, aligning with the individualized, retention-aware pricing logic proposed by Li et al. [6] for usage-based insurance. In practice, this reallocation could be implemented gradually - for example, by first applying the risk score to new policy underwriting before extending it to the existing book of business - so that the finance and actuarial teams can monitor realized claim experience against the model's predictions before committing the full reserve shift. Digital, dashboard-based monitoring of this kind has been shown to support timely operational decision-making in other public-sector reporting contexts, such as the field-officer distribution monitoring system described by Pamungkas and Rakhmi [22], and a similar monitoring layer could track realized-versus-predicted claim rates per risk tier over time.

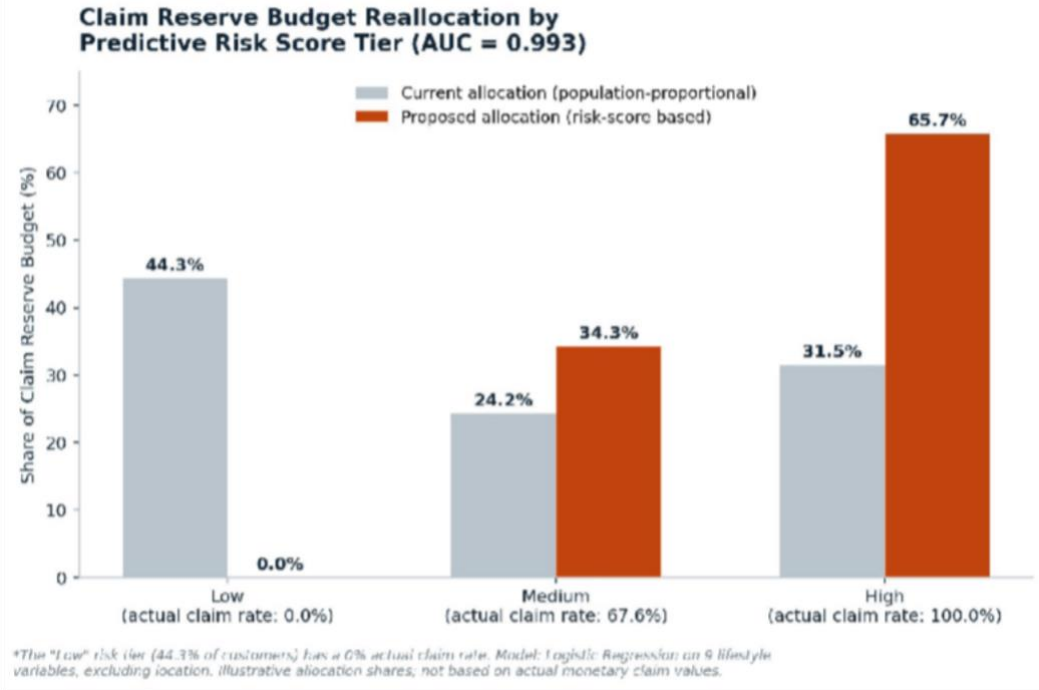


Figure 4. Simulated claim reserve budget reallocation by risk tier

3.6 Limitations

This study has six limitations, the first of which is foundational to how all other results should be read. First and most importantly, the dataset is synthetic and sourced from a public Kaggle repository rather than an operational insurer; Rankin et al. [17] found that models trained on synthetic health care data can systematically differ in accuracy from models trained on real data, particularly for tree-based classifiers, and Uddin and Lu [18] showed that meta-level statistical properties of Kaggle-sourced tabular datasets can materially affect ML performance in ways unrelated to the underlying real-world phenomenon. The near-identical, near-perfect AUC obtained from two structurally different models (Section 3.4) is best interpreted through this lens: as a signature of the data-generating rule rather than a discovery about real customer risk. Second, the dataset lacks monetary claim values, so the reallocation simulation reflects proportional shares rather than audited financial figures; production deployment requires linking risk scores to actual historical claim costs. Third, the claim-outcome variable's exact business definition could not be independently verified, and the reversed associations found for smoking and exercise frequency make this validation non-optional rather than precautionary. Fourth, the cross-sectional dataset precludes analysis of risk trends over time or of how customers migrate between risk tiers as behavior changes. Fifth, the near-deterministic separation between risk tiers (Section 3.3-3.4) means performance on live operational data is very likely to be substantially more modest than AUC = 0.993. Sixth, external validation on real-world, audited insurance data, following review protocols similar to those established for private health insurance analytics, is a precondition - not merely a recommendation - before any of this study's numerical findings inform actual budgeting decisions.

4. Conclusion

This study statistically tested the relative importance of geographic location versus individual lifestyle factors in explaining claim outcomes in a public, synthetic health-insurance dataset of 30,000 customers across 11 districts of Bekasi City, Indonesia, and used the comparison to demonstrate a replicable analytical framework for risk-based claim reserve budgeting. Location was statistically insignificant (Cramer's V = 0.019, p = 0.366), while both a logistic regression and a random forest model using nine lifestyle variables achieved an identical held-out test AUC of 0.993, separating customers into low-, medium-, and high-risk segments with claim rates of 0.0%, 67.6%, and 100.0%. Consistent with the objective stated in the Introduction, these findings illustrate how a claim reserve budgeting scheme could shift from geography-based to individual risk-score-based allocation. At the same time, the near-perfect, model-agnostic discrimination and the reversed directions found for smoking status, exercise frequency, and supplementary insurance are themselves evidence that this dataset's claim label follows a synthetic generation rule rather than real-world claims behavior; these results should therefore be read as a methodological proof of concept rather than as an actuarial finding ready for direct industry application. The essential next step before any operational use is external validation on real, audited insurance data, together with confirmation of the outcome variable's business definition by underwriting and claims stakeholders. Future work should incorporate real monetary claim data, external validation on independently audited operational datasets, additional ensemble architectures such as XGBoost [14] and multi-criteria decision-support methods such as OCRA/ROC [7] as a complementary ranking approach, instance-level explainability methods such as SHAP [23], broader model families as surveyed in standard statistical learning references [16], [25], and longitudinal data to capture risk trends over time.

References

[1] K. Kaushik, A. Bhardwaj, A. D. Dwivedi, and R. Singh, "Machine Learning-Based Regression Framework to Predict Health Insurance Premiums," Int. J. Environ. Res. Public Health, vol. 19, no. 13, art. 7898, 2022, doi: 10.3390/ijerph19137898.

[2] A. Alam and V. R. Prybutok, "Use of responsible artificial intelligence to predict health insurance claims in the USA using machine learning algorithms," Explor. Digit. Health Technol., vol. 2, pp. 30-45, 2024, doi: 10.37349/edht.2024.00009.

[3] M. Johnson, A. Albizri, and A. Harfouche, "Responsible artificial intelligence in healthcare: predicting and preventing insurance claim denials for economic and social wellbeing," Inf. Syst. Front., 2021, doi: 10.1007/s10796-021-10137-5.

[4] U. Orji and E. Ukwandu, "Machine learning for an explainable cost prediction of medical insurance," arXiv preprint, arXiv:2311.14139, 2023.

[5] W.-C. Lo, T.-H. Hu, C.-Y. Shih, H.-H. Lin, and J.-S. Hwang, "Impact of healthy lifestyle factors on life expectancy and lifetime health care expenditure: nationwide cohort study," JMIR Public Health Surveill., 2024, PMID: 39018094.

[6] H.-J. Li, X.-G. Luo, Z.-L. Zhang, S.-W. Huang, and W. Jiang, "A usage-based insurance (UBI) pricing model considering customer retention," Insurance Math. Econ., 2025, in press, doi: 10.1016/j.insmatheco.2025.06.006.

[7] M. Mayadi, R. W. P. Pamungkas, A. Azlan, K. Khairunnisa, and F. T. Waruwu, "Analisis Sistem Pendukung Keputusan Penentuan Kasi Terbaik Menerapkan Metode OCRA dengan Pembobotan Rank Order Centroid (ROC)," Build. Informatics Technol. Sci. (BITS), vol. 3, no. 3, pp. 393-399, 2021, doi: 10.47065/bits.v3i3.1100.

[8] B. Langenberger, T. Schulte, and O. Groene, "The application of machine learning to predict high-cost patients: a performance-comparison of different models using healthcare claims data," PLOS ONE, vol. 18, no. 1, art. e0279540, 2023, doi: 10.1371/journal.pone.0279540.

[9] M. Denuit, A. Charpentier, and J. Trufin, "Autocalibration and Tweedie-dominance for insurance pricing with machine learning," Insurance Math. Econ., vol. 101, pp. 485-497, 2021, doi: 10.1016/j.insmatheco.2021.08.010.

[10] S. Devriendt, K. Antonio, T. Reynkens, and R. Verbelen, "Sparse regression with multi-type regularized feature modeling," Insurance Math. Econ., vol. 96, pp. 248-261, 2021, doi: 10.1016/j.insmatheco.2020.11.010.

[11] H. Cramér, Mathematical Methods of Statistics. Princeton, NJ, USA: Princeton Univ. Press, 1946.

[12] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.

[13] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5-32, 2001, doi: 10.1023/A:1010933404324.

[14] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2016, pp. 785-794, doi: 10.1145/2939672.2939785.

[15] A. Agresti, Statistical Methods for the Social Sciences, 5th ed. Boston, MA, USA: Pearson, 2018.

[16] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009.

[17] D. Rankin, M. Black, R. Bond, J. Wallace, M. Mulvenna, and G. Epelde, "Reliability of supervised machine learning using synthetic data in health care: model to preserve privacy for data sharing," JMIR Med. Inform., vol. 8, no. 7, art. e18910, 2020, doi: 10.2196/18910.

[18] S. Uddin and H. Lu, "Dataset meta-level and statistical features affect machine learning performance," Sci. Rep., vol. 14, art. 1670, 2024, doi: 10.1038/s41598-024-51825-x.

[19] M. Michaelides, M. Pigeon, and H. Cossette, "Individual claims reserving using activation patterns," Eur. Actuarial J., 2023, doi: 10.1007/s13385-023-00355-3.

[20] S. Shaham, G. Ghinita, and C. Shahabi, "Models and mechanisms for spatial data fairness," Proc. VLDB Endow., vol. 16, no. 4, pp. 1014-1027, 2022, doi: 10.14778/3565816.3565820.

[21] B. Wijaya and R. W. P. Pamungkas, "IT governance and capacity building for smart city development," J. Int. Multidiscip. Res., vol. 3, no. 1, pp. 268-275, 2025.

[22] R. W. P. Pamungkas and K. Rakhmi, "Monitoring sebaran petugas lapangan Citarum Harum dalam pelaporan digital," J. Ilmiah Sist. Inf. dan Tek. Inform., vol. 9, no. 1, pp. 71-81, 2019.

[23] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 30, 2017, pp. 4765-4774.

[24] K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," Philos. Mag., Ser. 5, vol. 50, no. 302, pp. 157-175, 1900, doi: 10.1080/14786440009463897.

[25] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning: with Applications in R, 2nd ed. New York, NY, USA: Springer, 2021.