Comparing machine learning and binary regression approach for motor insurance prediction

Ridha Sefina Samosir[1], Jorge Luis Bazán Guzmán[2], Giselle Halim[1]
[1]Faculty of Computer Science and Design, Kalbis University, Jakarta, Indonesia
[2]Department of Applied Mathematics and Statistics, University of São Paulo, São Paulo, Brazil

Article Info

Article history:

Received Feb 6, 2025
Revised Oct 23, 2025
Accepted Nov 16, 2025

Keywords:

Binary regression
Imbalanced data
Insurance
Machine learning
Random forest

ABSTRACT

This study compares the performance of binary regression with the power cauchit (PC) link function and random forest in predicting motor insurance policyholder behavior using an imbalanced dataset. The dataset comprises 4,000 policyholders, with the response variable indicating whether a client purchased a full coverage plan (1) or not (0). Predictors include characteristics such as men, urban, private, age, and seniority. Binary regression was implemented using PyStan, while random forest was created with scikit-learn without additional hyperparameter tuning. Results demonstrate that random forest outperformed binary regression in a range of performance metrics, as well as specialized metrics suitable for imbalanced data. Findings point to the effectiveness of machine learning (ML) algorithms, exemplified by random forest, offer more robust performance in handling complex, imbalanced datasets compared to traditional statistical models. This highlights the potential of random forest to improve predictive accuracy in applications such as motor insurance policyholder behavior analysis.

Corresponding Author:

Ridha Sefina Samosir
Faculty of Computer Science and Design, Kalbis University
13210 Jakarta Timur, Jakarta, Indonesia
Email: ridha.samosir@kalbis.ac.id

1. INTRODUCTION

In the insurance industry, accurate prediction of customer behavior is critical for risk management, pricing strategies, and overall business decision-making. Understanding which policyholders are likely to purchase full coverage plans or other insurance products allows companies to tailor their offerings and optimize their marketing strategies. Traditionally, the financial sector, including insurance companies, has relied heavily on statistical modeling techniques such as binary regression to make predictions. Binary regression models have been a staple in financial analytics attributed to their interpretability and competence in modeling binary outcome variables effectively. Research has demonstrated that binary regression can be successfully applied in various financial contexts, including credit scoring and credit card fraud detection [1], [2]. While binary regression is reliable for capturing straightforward relationships and provides clear interpretability, challenges may arise when dealing with highly imbalanced datasets or more complex patterns in the data.

Binary regression, particularly with specialized link functions like the power cauchit (PC) link, has proven useful in modeling binary outcomes. The PC link function can provide more flexibility in certain data structures, allowing it to capture relationships that are not adequately represented by more conventional link functions. Previous research by [3] has demonstrated that a binary regression model with a PC link function (PC link) provided good results to predict full coverage purchases. This link function is especially advantageous in cases of imbalanced data due to its asymmetrical nature, which better handles the uneven distribution between classes [4]. Despite these strengths, binary regression, even with the PC link, may still face limitations when applied to datasets with highly nonlinear or intricate patterns. This underlines the need to explore alternative methods that can address these challenges effectively.

This challenge of data imbalance is pervasive in many real-world datasets, including those used in the insurance industry. Data imbalance takes place when the distribution of observations across classes is highly uneven, with one class greatly outnumbering the other, making it difficult for standard predictive models to identify the minority class accurately. This imbalance can skew the results of statistical models, causing them to favor the majority class and overlook the nuances within the minority class [5]. For insurance companies, this could mean missing out on identifying key segments of customers who are likely to purchase additional coverage, thereby impacting revenue and risk assessment strategies.

Advancements in machine learning (ML) have yielded techniques that address intricate and high-dimensional data more effectively compared to traditional statistical models. One such technique is the random forest algorithm is an ensemble approach that aggregates multiple decision trees to achieve higher predictive accuracy and better generalization [6]. The use of ML in insurance has shown promising results, particularly in areas like risk prediction, claim forecasting, and fraud detection [7]. The findings indicate that ML techniques, particularly random forests, can achieve good results in predicting outcomes in the insurance industry, therefore underscoring their potential to strengthen decision-making and streamline operations [8].

Comparable studies have examined how ML techniques are applied within diverse financial settings. For instance, research [9] and [10] has shown that ML models, like random forest, are able to bring good results in predicting insurance uptake. These studies have demonstrated that ML models can provide higher predictive accuracy and better generalization to unseen data, particularly in cases where the data distribution exhibits skewness or structural complexity. Additionally, another study focused on predicting loan behavior using various ML techniques revealed that random forest outperformed other models in terms of predictive accuracy [11], [12]. Supporting this trend, a study using data from a large automotive company in Brazil investigated the use of ML approaches, such as the random forest algorithm, to predict auto insurance claims. The results indicated that random forest outperformed other models, including logistic regression and Naïve Bayes, achieving accuracy, Kappa, and area under the curve (AUC) values of 0.8677, 0.7117, and 0.840, respectively [13]. These studies support the hypothesis that ML models, due to their flexibility and adaptability, have the potential to handle the intricacies of insurance data better than traditional binary regression models. This expanding body of literature highlights the growing relevance of employing ML techniques to enhance predictive modeling in finance.

This study compares the performance of binary regression with the PC link and the random forest algorithm in predicting motor insurance policyholders’ likelihood of purchasing full coverage. Using an imbalanced dataset, it evaluates which model achieves stronger predictive performance. While binary regression offers interpretability, random forest’s ML framework demonstrates greater adaptability to data complexity and imbalance. The findings highlight the growing potential of ML to enhance predictive accuracy in the insurance sector, emphasizing its role in advancing modern financial analytics beyond traditional statistical models.

2. METHOD

The research framework was structured to systematically fulfill the study’s objectives and ensure a structured approach to achieving the desired outcomes, as illustrated in Figure 1. The process began with problem identification, where the challenges of predicting motor insurance policy purchases in an imbalanced dataset were outlined. This was followed by a comprehensive literature study to examine existing approaches and pinpoint research gaps, which subsequently informed the formulation of the research question.

The next step involved data acquisition, where the secondary dataset of motor insurance policyholders was obtained and prepared for analysis. To ensure a fair comparison and eliminate potential biases, both the binary regression model employing the PC link and the random forest model were trained using the same data, without applying additional preprocessing steps. Furthermore, no hyperparameter tuning or adjustments were made to either model, ensuring that their default configurations were used for evaluation.

Afterward, model evaluation was conducted using several performance indicators, including accuracy, true positive rate (TPR), true negative rate (TNR), AUC- receiver operating characteristic (ROC), and several others, which will be further explained in the next section. Finally, conclusions were drawn based on the findings, offering insights into model performance and their broader implications. The section concludes with a synthesis of key findings and potential avenues for future exploration.

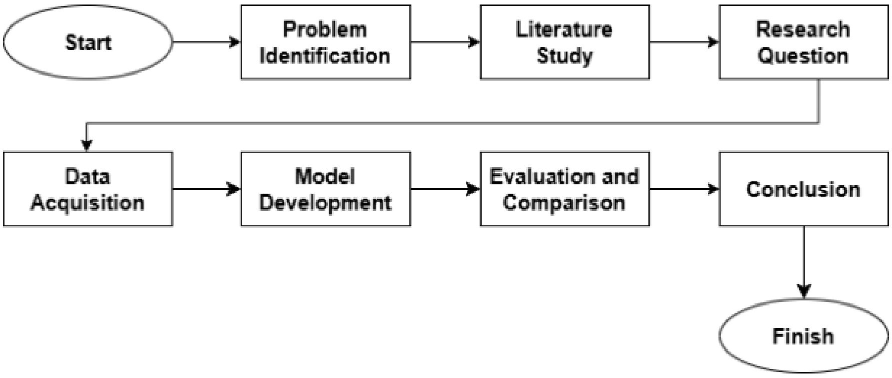

Figure 1. Research flow diagram

2.1. Motor insurance policyholders

Motor insurance policies are financial products designed to provide coverage against losses arising from vehicle-related incidents. They serve as a safety net for policyholders, protecting them from financial burdens associated with accidents, theft, or damage to their vehicles. Motor insurance policyholders play a critical role in the insurance ecosystem, as their experiences and perceptions significantly influence the dynamics between insurers and clients [14]. One type of motor insurance is full coverage insurance. This covers the costs of vehicle damage resulting from accidents, theft, vandalism, natural disasters, or impact with inanimate objects. Factors like age, income, and driving history influence the decision to purchase this type of insurance [3].

The use of customer demographic data allows insurers to refine risk assessments based on individual driving behaviors and their chance of purchasing a full coverage plan. For example, younger drivers or those with less driving experience may perceive a higher risk and thus opt for more comprehensive policies to mitigate potential financial repercussions.

2.2. Motor insurance policyholders' data

The dataset used in this study comprises secondary data sourced from the book “Predictive modeling applications in actuarial science” [15]. It includes information from 4,000 motor insurance policyholders, organized into six columns that capture client characteristics and their decision to purchase a full coverage motor insurance plan. The target variable (y) is binary, where 1 indicates the purchase of a policy (success) and 0 indicates no purchase (failure). This dataset is notably imbalanced, with only 34.7% of instances representing the positive class (y=1), posing a challenge in accurately predicting the minority class.

The predictors include both categorical and numerical features: MEN, a binary variable representing gender (0 for female and 1 for male); URBAN, a binary variable indicating the driving area (1 for urban areas and 0 for rural areas); PRIVATE, a binary variable signifying vehicle ownership status (1 for private vehicles and 0 for non-private vehicles); AGE, a numerical variable capturing the age of the policyholder; and SENIORITY, a numerical variable representing the length of employment in the company. The features were selected to reflect attributes likely influencing a client’s decision to purchase a full coverage plan.

2.3. Binary regression

Binary regression represents a statistical method designed to describe the relationship between a dichotomous dependent variable and one or more predictors. This method is widely used in clinical medicine and other fields to estimate the probability of a binary response variable based on a group of explanatory variables. The logistic regression model is particularly useful when the dependent variable is dichotomous, such as 0/1 [16].

A notable strength of binary regression lies in its capacity to model non-linear relationships between predictors and the dependent variable. The logistic function inherently captures non-linear effects, making it a robust choice for modeling complex relationships. Additionally, binary regression can handle both continuous and categorical predictor variables, making it versatile for a wide range of applications [17].

2.4. Binary regression with power cauchit link

The PC link function is a type of asymmetric link function that can be used in binary regression models. This is an extension of the cauchit link, which provides flexibility in modeling data with heavy tails or imbalances. This allows the link function to adjust the tails of the distribution to better handle outliers and imbalanced data. By adjusting the power parameter, the PC link function can mitigate the impact of the imbalance, ensuring that the model is sensitive enough to predict the minority class without being overwhelmed by the majority class [18].

Markov chain monte carlo (MCMC) algorithms serve as robust computational methods for approximating the parameters of complex models, including binary regression models with the PC link function. MCMC methods allow for the simulation of randomly generating posterior samples for the model parameters, enabling Bayesian inference. This approach is particularly useful when dealing with identifiability issues that can arise with skewed link functions, such as the PC link [19]. In practice, implementing binary regression with the PC link can be done using statistical software packages such as Python libraries like PyStan.

2.5. Machine learning

ML, a branch of artificial intelligence, focuses on designing algorithms and statistical models that allow systems to perform specific tasks independently. The models derive knowledge from data by recognizing underlying patterns and making data-driven decisions based on them. The fundamental concept is to allow machines to independently learn from experience, progressively improving their accuracy as more data becomes available. This process typically involves training a model on a dataset, which it uses to make predictions when presented with new data [20].

ML methods are commonly classified into several paradigms, such as supervised learning, where models are fitted to labeled data; unsupervised learning, which analyzes unlabeled data to discover hidden relationships; semi-supervised learning, where the model combines elements of both supervised and unsupervised learning; and reinforcement learning, in which an agent learns optimal actions through rewards and penalties based on performance. The applications of ML are vast, spanning fields such as healthcare, finance, and many others [21].

2.6. Random forest

In ML, models such as random forest generate numerous decision trees during training and integrate their results by majority vote in classification or by averaging in regression to produce stable predictions. This ensemble strategy enhances robustness, effectively manages complex, high-dimensional data, and mitigates the risk of overfitting [6].

Random forest operates by creating a multitude of decision trees, each trained on a different subset of the dataset. This subset is created through bootstrapping, a technique that samples the data with replacement, ensuring diversity among the trees. For each tree, a random subset of features is selected at each split point, which prevents any single feature from dominating the model and enhances the model's ability to generalize [22].

The primary advantage of random forest lies in its ability to handle non-linear relationships and interactions between variables without the need for extensive parameter tuning. Furthermore, it is inherently capable of dealing with imbalanced datasets. Random forest also conducts feature selection to pinpoint the most significant features within the data, enhancing the model's performance on datasets. Other benefits of random forest include its resilience to outliers, enhanced predictive accuracy, and its ability to manage overfitting [23]. This study implements random forest algorithm to predict whether someone will buy the motor insurance policy using an imbalanced dataset, where the challenge lies in accurately predicting the minority class.

2.7. Evaluation metrics

The models' performance can be seen in the values of true negative (TN), false positive (FP), false negative (FN), and true positive (TP) rates. TN represents correct predictions of non-purchasers, while FP denotes incorrect predictions of purchasers. FN indicates failure to identify actual purchasers, and TP signifies correct identification of purchasers. These four outcomes are essential for evaluating the performance of the models, as they provide the basis for calculating key metrics, which help determine the model's overall predictive power and reliability.

The performance of both models was evaluated using common ML metrics: accuracy (ACC), precision, recall (sensitivity or TPR), specificity (TNR), F1-score, and area under the ROC curve (AUC) [24], defined respectively by,

$$ACC = \frac{TP+TN}{TP+FP+FN+TN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$TPR = \frac{TP}{TP+FN} \quad (3)$$

$$TNR = \frac{TN}{TN+FP} \quad (4)$$

$$AUC - ROC = \int_0^1 \quad TPR(t)dt \quad (5)$$

$$F1 - score = 2 * (Precision * Recall)/(Precision + Recall) \quad (6)$$

These metrics were employed to evaluate each model's predictive capability. Accuracy indicates the overall proportion of correct classifications but can be misleading in imbalanced datasets dominated by one class. Precision reflects the model's ability to reduce false positives, while TPR (recall) measures the proportion of actual positives correctly identified. TNR (specificity) assesses the model's accuracy in recognizing negatives. The F1 score, defined as the harmonic means of precision and recall, offers a balanced assessment particularly suitable for imbalanced data.

The AUC quantifies the area beneath the ROC curve, illustrating the trade-off between TPR and FPR across various thresholds. A higher AUC value indicates superior discriminative power, reflecting the model's ability to distinguish clients who are likely to purchase full coverage from those who are not [25].

In addition to the standard metrics, we employed evaluation measures specifically designed for imbalanced datasets [26]. Additional evaluation metrics such as the critical success index (CSI), Gilbert skill score (GSS), symmetric gini score (SGS), Sokal and Sneath index (SSI), faith index (FAITH), Matthews correlation coefficient (MCC), Geometric mean (G_M), and Cohen's Kappa (KAPPA) were also utilized for model comparison. The definitions of these measures are presented as follows:

$$CSI = \frac{TP}{TP+FP+FN} \quad (7)$$

$$GS = \frac{TP * TN - FP * FN}{(FN+FP) * (TP+FP+FN+TN) \ + (TP*TN-FP*FN)} \quad (8)$$

$$SGS = \frac{3 * GS + 1}{4} \quad (9)$$

$$SSI = \frac{TP}{(TP + 2 * FP + 2 * FN)} \quad (10)$$

$$FAITH = \frac{TP+0.5*TN}{TP+TN+FP+FN} \quad (11)$$

$$MCC = \frac{TP+TN-FP*FN}{\sqrt{}((TP + FP) * (TP + FN) * (TN + FP) * (TN + FN))} \quad (12)$$

$$G\_M = \sqrt{}(TPR * TNR) \quad (13)$$

These metrics are designed to offer a nuanced assessment of model performance beyond simple accuracy, which can be misleading in the presence of imbalanced data. As with any performance metric, higher values in these metrics indicate better model performance.

3. RESULTS AND DISCUSSION

For this research, we apply the binary regression with the PC link function to a dataset of 4,000 motor insurance policyholders available in [15], where the target variable (y) indicates whether a client purchased a full coverage plan (1 for success and 0 for failure). Given the imbalance in the data, where instances of 0 vastly outnumber instances of 1 (34.7% proportion of ones), the PC link function is expected to provide superior model performance by accurately predicting the minority class while maintaining overall model stability based on previous study [3]. The effectiveness of this method will be evaluated against random forest, a ML algorithm, to determine which approach yields better predictive accuracy for this imbalanced dataset.

To assess and compare the predictive performance of binary regression with the PC link function and random forest, we used all 4,000 data points for training and testing. The potential predictors included individual characteristics such as men, urban, private, age, and seniority. The binary regression model was constructed using PyStan, a Python interface for stan, which allows for Bayesian inference and advanced statistical modeling. Additionally, MCMC was used to generate posterior distributions for the model parameters, providing uncertainty estimates. The estimated model coefficients and their associated uncertainties were obtained by summarizing the posterior samples. Random forest was implemented using scikit-learn, a widely used Python toolkit for ML, without additional hyperparameter tuning. Table 1 displays a comparative evaluation between the binary regression model with PC link and the random forest model, focusing on their classification outcomes. Table 2 summarizes the performance of both models across several key evaluation measures, such as accuracy, precision, F1-score, TPR, and TNR.

Table 1. Confusion matrix comparison of binary regression and random forest

Model	TN	FP	FN	TP
Binary regression (PC link)	1765	848	211	1176
Random forest	2397	216	250	1137

Table 2. Metrics comparison of binary regression and random forest

Model	Accuracy	Precision	F1-Score	AUC	TPR	TNR
Binary regression (PC link)	0.74	0.58	0.69	0.79	0.85	0.68
Random forest	0.88	0.84	0.83	0.87	0.82	0.92

Random forest outperformed binary regression with the PC link function across nearly all performance metrics, except for TPR. The higher TPR for binary regression suggests that it is more effective at identifying clients who bought full coverage, meaning it is better at catching the minority class (those who purchased the plan) even in the imbalanced data scenario. However, this increased sensitivity to the positive class likely came at the expense of specificity, as the model struggled to correctly classify the negative class, resulting in a lower TNR. This trade-off between TPR and TNR is one of the typical difficulties in handling imbalanced datasets that improving one metric often results in the deterioration of another. However, the Random Forest model demonstrates superior performance in terms of specificity (TNR), where a higher value reflects a stronger capability to correctly identify negative instances. In this context, a higher specificity indicates that the model demonstrates stronger capability in accurately detecting clients who did not buy the full coverage plan (i.e., those coded as 0).

The fact that random forest performed better in terms of specificity suggests that it was more cautious in classifying observations as positive (purchasing the full coverage), leading to fewer false positives. This conservative approach means the model is less likely to incorrectly predict a client will buy full coverage when they will not, which is beneficial in scenarios where avoiding false positives is critical (e.g., assuming more clients will buy full coverage than actually do). Related to this, the higher AUC value for the random forest model can be attributed to the model's much higher TNR. The AUC score is derived by plotting the TPR against the FPR (false positive rate, which is 1 - TNR) across various threshold levels [25]. A higher AUC score indicates that the model has a better balance between TPR and TNR, meaning it can effectively distinguish between positive and negative cases.

Further evaluation is presented in Table 3, which compares additional performance metrics such as critical success index (CSI), Gilbert skill score (GS), symmetric gini score (SGS), Sokal and Sneath index (SSI), faith index, Matthews correlation coefficient (MCC), geometric mean (G_M), and Cohen’s Kappa. The results presented in the previous tables indicate that random forest surpasses binary regression in nearly all performance indicators, except for the TPR. This finding emphasizes random forest’s advantage in managing imbalanced data and generating more reliable predictions compared to conventional statistical methods.

Table 3. Advanced evaluation metrics comparison of binary regression and random forest

Model	CSI	GS	SGS	SSI	FAITH	MCC	G_M	KAPPA
Binary regression (PC link)	0.53	0.53	0.48	0.36	0.51	0.5	0.76	0.47
Random forest	0.71	0.59	0.69	0.55	0.58	0.74	0.87	0.74

Extensive studies have compared random forest with traditional logistic regression, yet limited research has examined its comparison with binary regression, which has been modified using a Bayesian approach, particularly with custom link functions like PC. This study aims to fill this gap by evaluating the performance of random forest and Bayesian binary regression on an imbalanced motor insurance dataset. Prior studies have highlighted random forest's robustness in handling complex, non-linear relationships and imbalanced datasets, where traditional logistic regression often struggles due to its reliance on linear assumptions [27]. This makes random forest a suitable candidate for datasets with intricate structures, as seen in this study.

Although binary regression improves upon traditional logistic regression by incorporating Bayesian priors and flexible link functions, such as the PC link used in this study, it remains fundamentally rooted in the statistical assumptions of logistic regression [4]. Therefore, insights derived from random forest vs logistic regression comparisons provide a meaningful context for evaluating random forest against Bayesian binary regression. Consistent with existing literature on random forest's superiority over logistic regression [28], this study demonstrates that random forest outperforms Bayesian binary regression across most evaluation metrics. These findings highlight the random forest’s strength in managing class imbalance and capturing complex non-linear relationships within the dataset.

Logistic regression often serves as a baseline for comparison in classification problems; however, it is important to recognize that this study evaluates random forest against Bayesian binary regression with a flexible link function. Unlike standard logistic regression, which struggles with capturing non-linearity, Bayesian binary regression with the PC link function introduces greater flexibility in modeling complex relationships. The key question in this study is whether the PC link function can sufficiently address the challenges posed by the data imbalance and non-linearity to perform better than random forest, a ML approach specifically designed to handle such complexities.

Interestingly, Bayesian binary regression achieved a higher TPR, suggesting that its flexibility in modeling the response variable allows it to capture more positive instances. Despite that, it struggled to maintain specificity, leading to higher false positives. This may suggest that overly flexible link functions, such as PC, could increase the risk of overfitting to certain patterns in the data [29]. In contrast, random forest's ensemble approach provided a more balanced trade-off between TPR (sensitivity) and TNR (specificity), resulting in superior performance across most evaluation metrics even with imbalanced datasets.

The results validate random forest's effectiveness in imbalanced binary classification, demonstrating its superiority over traditional statistical methods, even with Bayesian enhancements. By combining insights from ML and Bayesian statistics, this study bridges gaps in literature with a comprehensive evaluation framework. Additionally, the findings reinforce the potential of ML models like random forest to enhance predictive accuracy in real-world applications. This highlights their role in improving binary classification and decision-making processes, particularly in the financial sector.

4. CONCLUSION

This study conducted a comparative analysis of the performance between binary logistic regression with the PC link function and the random forest algorithm in predicting motor insurance policyholder behavior using an imbalanced dataset. The response variable indicated whether a policyholder opted for a full coverage plan, with an imbalance favoring non-purchasers. The predictors included key individual characteristics such as MEN, URBAN, PRIVATE, AGE, and SENIORITY.

The findings indicate that the random forest model consistently outperformed the binary regression model across almost all evaluated metrics. Random forest’s superior performance, particularly in handling the minority class, underscores its robustness in dealing with imbalanced data, providing more accurate and reliable predictions in this context. This demonstrates the potential of ML models like random forest to surpass traditional statistical methods like binary regression, particularly in applications involving complex, imbalanced datasets. This study suggests that random forest should be considered a strong candidate for predictive modeling in similar contexts, where accurate identification of minority class events is crucial.

These findings also emphasize the importance of further research into the potential limitations of binary regression with custom link functions when applied to imbalanced datasets. While the PC link function provides more flexibility than standard logistic regression, its effectiveness in handling extreme class imbalances remains a key area for investigation.

Future studies could explore hybrid approaches that integrate the adaptability of Bayesian methods with the robustness of ensemble models like random forest, potentially enhancing predictive performance and model stability in financial and insurance applications. Additionally, subsequent studies may explore how hyperparameter optimization influences the performance of random forest and other ML algorithms, as well as the application of advanced techniques to further improve performance in highly imbalanced scenarios.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the anonymous reviewers for their thoughtful insights and meticulous feedback, which played a vital role in refining and elevating the overall quality of this work. We would also like to acknowledge the contributions of the open-source community whose libraries and tools were instrumental in conducting this research. Furthermore, we extend our heartfelt thanks to Kalbis University and University of São Paulo for their invaluable support and encouragement throughout the course of this study.

FUNDING INFORMATION

This work was carried out independently, with no financial backing from governmental, commercial, or non-profit organizations. All stages of the study, including simulation and analysis, were conducted without external financial support. The publication of this paper is personally funded by the corresponding author.

AUTHOR CONTRIBUTIONS STATEMENT

In accordance with the Contributor Roles Taxonomy (CRediT), the specific contributions of each author are as follows. Ridha Sefina Samosir (corresponding author) conceptualized the study, designed the research methodology, and conducted the initial software implementation. She also prepared the original draft, reviewed and edited the manuscript, managed the project administration and funding, and supervised the overall research process. Jorge Luis Bazán Guzmán contributed to the conceptualization and methodology, provided the dataset used in this study, and supplied key resources and references supporting the binary regression simulation. He also participated in the manuscript review, supervision, and project administration. Giselle Halim performed the simulation of both models, conducted model evaluation and comparative analysis, prepared data visualization, and contributed to the original draft writing and manuscript editing. All authors contributed to the discussion of results, reviewed the final version of the manuscript, and approved its submission.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Ridha Sefina Samosir	✓	✓	✓	✓					✓	✓		✓	✓	✓
Jorge Luis Bazán Guzmán	✓	✓		✓		✓	✓	✓		✓		✓	✓	
Giselle Halim			✓	✓	✓				✓	✓	✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis

I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting

Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper. The authors state no conflict of interest.

DATA AVAILABILITY

The dataset that supports the findings of this study is publicly available as secondary data from Predictive Modeling Applications in Actuarial Science published by Cambridge University Press at http://doi.org/10.1017/cbo9781139342674.001. The data, referred to as "FullCoverage," can be accessed at https://instruction.bus.wisc.edu/jfrees/jfreesbooks/PredictiveModelingVol1/predictive-modeling-foundations/chapter-3.html.

REFERENCES

[1] S. Jha, M. Guillen, and J. Christopher Westland, "Employing transaction aggregation strategy to detect credit card fraud," *Expert Systems with Applications*, vol. 39, no. 16, pp. 12650–12657, Nov. 2012, doi: 10.1016/j.eswa.2012.05.018.

[2] V. L. Miguéis, D. F. Benoit, and D. Van den Poel, "Enhanced decision support in credit scoring using Bayesian binary quantile regression," *Journal of the Operational Research Society*, vol. 64, no. 9, pp. 1374–1383, Sep. 2013, doi: 10.1057/jors.2012.116.

[3] J. L. Bazán, F. Torres-Avilés, A. K. Suzuki, and F. Louzada, "Power and reversal power links for binary regressions: an application for motor insurance policyholders," *Applied Stochastic Models in Business and Industry*, vol. 33, no. 1, pp. 22–34, Jan. 2017, doi: 10.1002/asmb.2215.

[4] L. F. M. Reis, D. C. Nascimento, P. H. Ferreira, and F. Louzada, "Fixing imbalanced binary classification: an asymmetric Bayesian learning approach," *PLOS ONE*, vol. 19, no. 10, p. e0311246, Oct. 2024, doi: 10.1371/journal.pone.0311246.

[5] H. He and E. A. Garcia, “Learning from imbalanced data,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.

[6] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[7] T. Poufinas, P. Gogas, T. Papadimitriou, and E. Zaganidis, “Machine learning in forecasting motor insurance claims,” *Risks*, vol. 11, no. 9, p. 164, Sep. 2023, doi: 10.3390/risks11090164.

[8] S. T. Lim, J. Y. Yuan, K. W. Khaw, and X. Chew, “Predicting travel insurance purchases in an insurance firm through machine learning methods after COVID-19,” *Journal of Informatics and Web Engineering*, vol. 2, no. 2, pp. 43–58, Sep. 2023, doi: 10.33093/jiwe.2023.2.2.4.

[9] N. K. Yego, J. Kasozi, and J. Nkurunziza, “A comparative analysis of machine learning models for the prediction of insurance uptake in Kenya,” *Data*, vol. 6, no. 11, p. 116, Nov. 2021, doi: 10.3390/data6110116.

[10] A. S. Alshamsi, “Predicting car insurance policies using random forest,” in *2014 10th International Conference on Innovations in Information Technology (IIT)*, Nov. 2014, pp. 128–132, doi: 10.1109/INNOVATIONS.2014.6987575.

[11] M. Anand, A. Velu, and P. Whig, “Prediction of loan behaviour with machine learning models for secure banking,” *Journal of Computer Science and Engineering (JCSE)*, vol. 3, no. 1, pp. 1–13, Feb. 2022, doi: 10.36596/jcse.v3i1.237.

[12] M. Madaan, A. Kumar, C. Keshri, R. Jain, and P. Nagrath, “Loan default prediction using decision trees and random forest: a comparative study,” *IOP Conference Series: Materials Science and Engineering*, vol. 1022, no. 1, p. 012042, Jan. 2021, doi: 10.1088/1757-899X/1022/1/012042.

[13] M. Hanafy and R. Ming, “Machine learning approaches for auto insurance big data,” *Risks*, vol. 9, no. 2, p. 42, Feb. 2021, doi: 10.3390/risks9020042.

[14] S. S. Ajemunigbohun, T. O. Oluwaleye, and A. B. Sogunro, “Claims handling process attributes: perceptions of motor insurance policyholders in Lagos, Nigeria,” *Journal of Corporate Governance, Insurance, and Risk Management*, vol. 9, no. 1, pp. 136–154, Aug. 2022, doi: 10.51410/jcgirm.9.1.9.

[15] E. W. Frees, R. A. Derrig, and G. Meyers, “Predictive modeling in actuarial science,” in *Predictive Modeling Applications in Actuarial Science*, Cambridge University Press, 2014, pp. 1–10.

[16] J. R. Wilson and K. A. Lorenz, “Introduction to binary logistic regression,” in *ICSA Book Series in Statistics*, Springer, Cham, 2015, pp. 3–16.

[17] A. Agresti, *Categorical data analysis, 3rd edition*. Wiley, 2012.

[18] R. Galo, R. Marcelo Rossi, D. Corrêa Alves, and R. Rosseto de Oliveira, “Bayesian binary regression using power and power reverse link functions: an application to premature birth data,” *Brazilian Journal of Biometrics*, vol. 41, no. 2, pp. 131–143, Jun. 2023, doi: 10.28951/bjb.v41i2.604.

[19] J. V. B. de Freitas and C. L. N. Azevedo, “Regression models for binary data with scale mixtures of centered skew-normal link functions,” *arXiv preprint arXiv:2407.14748*, Jul. 2024, [Online]. Available: http://arxiv.org/abs/2407.14748.

[20] G. Rebala, A. Ravi, and S. Churiwala, “Machine learning definition and basics,” *An Introduction to Machine Learning*, pp. 1–17, 2019, doi: 10.1007/978-3-030-15729-6_1.

[21] R. Pugliese, S. Regondi, and R. Marini, “Machine learning-based approach: global trends, research directions, and regulatory standpoints,” *Data Science and Management*, vol. 4, pp. 19–29, Dec. 2021, doi: 10.1016/j.dsm.2021.12.002.

[22] “Random forest classifier documentation,” *Scikit-Learn Developers*. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (accessed Feb. 27, 2024).

[23] M. Schonlau and R. Y. Zou, “The random forest algorithm for statistical learning,” *The Stata Journal: Promoting communications on statistics and Stata*, vol. 20, no. 1, pp. 3–29, Mar. 2020, doi: 10.1177/1536867X20909688.

[24] J. D. Kelleher, B. Mac Namee, and A. D’Arcy, *Fundamentals of machine learning for predictive data analytics - algorithms, worked examples, and case studies*. London, England: The MIT Press Cambridge, Massachusetts, 2015.

[25] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.

[26] A. de la Cruz Huayanay, J. L. Bazán, V. G. Cancho, and D. K. Dey, “Performance of asymmetric links and correction methods for imbalanced data in binary regression,” *Journal of Statistical Computation and Simulation*, vol. 89, no. 9, pp. 1694–1714, Jun. 2019, doi: 10.1080/00949655.2019.1593984.

[27] D. Muchlinski, D. Siroky, J. He, and M. Kocher, “Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data,” *Political Analysis*, vol. 24, no. 1, pp. 87–103, 2016, doi: 10.1093/pan/mpv024.

[28] P. A. Sunarya, U. Rahardja, S. C. Chen, Y. M. Li, and M. Hardini, “Deciphering digital social dynamics: a comparative study of logistic regression and random forest in predicting e-commerce customer behavior,” *Journal of Applied Data Sciences*, vol. 5, no. 1, pp. 100–113, 2024, doi: 10.47738/jads.v5i1.155.

[29] D. M. Hawkins, “The problem of overfitting,” *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 1, pp. 1–12, Jan. 2004, doi: 10.1021/ci0342472.

BIOGRAPHIES OF AUTHORS

Ridha Sefina Samosir hold the doctoral, master, and bachelor’s in computer science from Sanata Dharma University, Indonesia University, and Bina Nusantara University. She is currently as an associate professor at Kalbis University in 2016. She used to teach business data analytics and big data analytics. Since 2011 to 2020 as the head of Study Program and 2020 – 2022 as a dean of computer science and design faculty. Her research interests are artificial intelligence, computer vision, and data mining. She has published many publications that are indexed by international databases such as Scopus. She got grant from Indonesia Government from her research about AI and Law. She can be contacted at email: ridha.samosir@kalbis.ac.id.

Jorge Luis Bazán Guzmán is associate professor in the Department of Applied Mathematics and Statistics at the Institute of Mathematical and Computer Sciences at the University of São Paulo, Brazil. He has a degree in statistical engineering from the National Agrarian University La Molina of Peru and a Ph.D. in statistics from the Institute of Mathematics and Statistics of the University of São Paulo, additionally, he did a postdoctoral period at the Department of Statistics of the University of Connecticut in the United States. His research primarily focuses on data science and statistics, encompassing wide array of topics such as regression and classification models, latent variable models (item response theory models, cognitive diagnostic models), Bayesian inference, categorical data, psychometrics and statistical education. He can be contacted at email jlbazan@icmc.usp.br.

Giselle Halim is affiliated with the Faculty of Computer Science and Design at Kalbis University, Jakarta. With a strong background in machine learning and its applications, she has contributed to research on integrating artificial intelligence into web applications for healthcare. Notably, a paper on the implementation of ML for cervical cancer risk detection in web applications has been presented at an international conference. Her research interests include predictive modeling, data-driven decision-making, and the intersection of artificial intelligence with real-world applications. She can be contacted at email: gisellehalim27@gmail.com.