

A Machine Learning-Based Analysis of Health Insurance Expenditures in Rwanda

Mr. Ephraim Uyisenga¹, Dr. Roger Muremyi²

¹Data Scientist, Insurance, Mmi

²Lecturer, African Center Of Excellence In Data Science, University Of Rwanda

Abstract

Background: Rising healthcare costs among elderly insured members present sustainability challenges for Rwanda's health insurance schemes. This study applies machine learning (ML) to predict healthcare expenditures for elderly members, identify key cost drivers, and develop a deployable actuarial decision-making tool.

Objective: To analyze and predict health expenditure patterns of elderly insured members using ML, compare predictive models, and provide policy-relevant insights.

Methods: Claims data (2022–2024) from the Military Medical Insurance (MMI) scheme were analyzed using Multiple Linear Regression, Random Forest, Gradient Boosting, and XGBoost. Data were split into training (80%) and testing (20%) sets. Models were evaluated using R^2 , MAE, and RMSE. SHAP values were used for interpretability.

Results: Ensemble models outperformed linear models; Random Forest achieved the highest accuracy ($R^2 = 0.95$ after tuning). Age, BMI, and service categories (Acts, imaging, and hospitalization) were major predictors. NCDs affected costs indirectly through increased use of high-cost services.

Conclusion: ML can improve risk adjustment, preventive targeting, and pricing strategies in Rwanda's health insurance. The deployed Streamlit tool provides practical real-time predictions for insurers.

Keywords: Machine Learning, Health Insurance, Elderly, Healthcare Expenditure, Rwanda, Actuarial Science

1. Introduction

Population aging is accelerating globally, with 80% of older adults projected to live in low- and middle-income countries by 2050 (Gerland, et al., 2022). In Rwanda, this shift, combined with rising non-communicable diseases (NCDs), increases health insurance costs and strains sustainability (Malene, Alexander, Seetha, & Jes, 2024). Elderly members often have higher healthcare utilization, particularly for costly services such as imaging and hospitalization.

Traditional actuarial models often rely on linear assumptions, limiting their ability to capture complex interactions among health determinants. This study addresses that gap by applying machine learning to predict health insurance expenditures among the elderly, aiming to improve cost forecasting and policy planning using machine learning that capture nonlinearity.

This study has main objective to analyze and predict the health expenditure patterns of the aged population enrolled in health insurance schemes in Rwanda using machine learning techniques. The main objective will be followed by specific objectives as to examine the demographic trends and healthcare utilization

patterns of the aged population in Rwanda, to apply supervised machine learning models for predicting future healthcare costs of aged insured members.

To achieve this objective, following research questions were asked: To what extent does age correlate with healthcare costs and claims frequency among insured individuals in Rwanda? Which machine learning method provides the most accurate and interpretable forecasts for elderly healthcare costs or claim likelihoods in the context of Rwanda? This study hypothesizes that the aged population in Rwanda has a significant effect on health insurance expenditures. Alternatively, age has no significant effect on health insurance expenditure.

This study applies interpretable ML methods to predict elderly health insurance expenditures and identify actionable cost drivers.

2. Literature review

2.1 Background

Prior studies have explored health insurance utilization and cost determinants globally, but few have applied ML in low-income contexts. Grossman's health capital model provides a theoretical foundation, linking health status to economic behavior (Grossman, 1972). This study builds on existing literature by integrating ML techniques with actuarial science to model expenditure patterns in Rwanda's elderly population.

2.2 Global context

Globally, the health of older adults is increasingly challenged by chronic, non-communicable diseases (NCDs) such as heart disease, cancer, and diabetes. As populations age, these conditions have become more prevalent, demanding a realignment of healthcare priorities and resources to safeguard seniors' well-being and quality of life (Malene, Alexander, Seetha, & Jes, 2024).

2.3 Regional context

In East Africa, demographic shifts mirror global trends, with the elderly population (60+ years) growing steadily due to declining fertility rates, urbanization, and migration. However, health insurance coverage for older adults remains limited. In Kenya, the National Hospital Insurance Fund (NHIF) primarily covers formal workers, with elderly enrollment constrained by the dominance of informal sector employment (Mugo, 2023). Uganda's health insurance penetration is below 5%, with no dedicated program for older adults (Kajubi, Tumushabe, & Turyareeba, 2021). In Tanzania, health financing is fragmented, and voluntary health insurance uptake among the elderly remains below 10% (Muhanga & Mubyazi, 2022).

2.4 Rwandan perspective

In Rwanda, Muremyi et al. (2023) analyzed the impact of health insurance on healthcare utilization using logistic regression and decision tree models. While the decision tree effectively classified districts based on healthcare usage similarities, the study excluded certain variables related to health expenditure and was limited to binary classification, leaving room for regression-based cost prediction approaches (Roger, Dominique, François, & Kabano, 2023).

Rwanda's Ministry of Local Government has acknowledged that rapid aging, combined with changing family structures, labor market trends, and migration, will soon impose substantial social, economic, and health demands on the country's institutions (MINALOC, 2021). In response, the government launched the *EjoHeza* program to support the elderly population, including government employees, private sector workers, and the self-employed (RSSB, 2024).

2.5 Model comparison studies

From a methodological perspective, (Langenberger, Schulte, & Groene, 2023) compared machine learning algorithms in predicting future high-cost patients using German sickness fund data over a three-year period, concluding that random forest was the best-performing algorithm (95% CI). Although the study used a classification approach with German data, there is potential to adapt such methods to the Rwandan context using regression models to forecast healthcare costs.

Furthermore, (Kaushik, Bhardwaj, Dwivedi, & Singh, 2022) emphasized that artificial intelligence (AI) and machine learning are transforming health insurance operations, enabling insurers to design policies more efficiently and deliver services faster, ultimately saving time and money for both insurers and policyholders.

2.6 Theoretical framework

The study is grounded in **Grossman's health capital model**, which conceptualizes health as a durable good influenced by individual choices and external factors. This framework supports the idea that health expenditures are shaped by both demographic and behavioral variables.

3. Methodology

3.1 Overview

Machine learning offers significant advantages compared to traditional statistical methods in analyzing data. In the context of this study, the focus will be ML models (Random Forests regression, Gradient boosting, and XGBoost), which may learn complex, non-linear patterns and interactions without requiring strict assumptions.

3.2 Data Source

The secondary data from MMI that comprise medical expenses of beneficiaries from 3 years 2022-2024 daily series covering demographics, health metrics, and service usage will be used. This data will help us on exploring the financial burden imposed by elderly people in health insurance.

3.3 Variables

Dependent: Daily claim amount.

Independent: Age, gender, BMI, NCD status, service category (Acts, imaging, hospitalization, consultation, labo), blood pressure, regular exercise.

3.4 Models

Multiple Linear Regression (baseline), Random Forest Regressor (RF) an ensemble learning method designed for classification, regression, and other predictive tasks. It works by training multiple decision trees and combining their outputs either through majority voting (classification) or averaging (regression) (Breiman, 2001). Gradient Boosting (GB) a machine learning method that creates strong predictive models by combining multiple weak learning models, specifically optimized boosted decision trees. Gradient boosting identifies the weaknesses of the weak learners using gradients in the loss function, which measures how well the model's coefficients fit the given data (DANGETI, 2017). XGBoost (XG) (eXtreme Gradient Boosting) is an advanced gradient boosting framework renowned for its efficiency, accuracy, and scalability (Chen & Guestrin, 2016).

3.5 Evaluation Metrics

R^2 , also known as the coefficient of determination. It is a popular way to tell how well your model fits the data. It is a measure of the percentage of the variation in your dependent variable that your independent variables can explain or predict (Jian, 2024).

Mean Absolute Error (MAE) Measures the average magnitude of prediction errors without considering direction, providing an easily interpretable indicator of accuracy (Wang & Lu, 2018).

Root Mean Square Error (RMSE) (Root Mean Squared Error) comes from taking the square root of the MSE. Hence, this operation doesn't change the comparative performance rankings of models; it produces a measure that shares the same units as the target variable (y). This makes the RMSE particularly useful (Hodson, 2022).

3.6 Parameter tuning

Parameter tuning (or hyperparameter optimization) is the process of selecting the best set of hyperparameters for a machine learning model to improve its performance. Unlike model parameters (learned during training), hyperparameters are set before training and control the learning process (Bergstra & Bengio, 2012). In this study Hyper parameter tuning with 5-fold cross-validation was applied. The benefit of that it improves model accuracy and generalization, Prevents overfitting or under fitting.

Model	Estimators	Max depth	Min sample split or Learning rate
Random Forest	100, 200, 300	None, 10, 20	2, 5, 10
XGBoost	100, 200, 300	3, 5, 7	0.01, 0.1, 0.2

Table 1 Parameters of tuning

4. Results

4.1 Descriptive Analysis

Elderly claims correlated strongly with age (51%) and service category (43%). Higher BMI tended to be associated with increased claims, with regular exercise mitigating some costs.

4.2 Model Performance

	Actual	LR	RF	GR	XG
4344	8968	17408.78	18116.29	17454.82	18993.36
12416	8596.8	77169.12	12670.55	18759	11145.08
8046	13452	65829.72	16445.26	22141.3	24670.35
10713	112547	172553.6	111775	107322.4	114419.3
15526	484837	219931.8	329818	437706.9	429216.7
...	...	...	...	...	...
2390	14254.6	-23386.6	12366.18	13459.77	12256.08
15172	375000	214694.5	352786.7	335993.1	353552.3
5916	7520	154700	37339.42	63924.72	65467.86
16258	7985.8	51596.37	11972.39	23507.11	11415.83
5924	8059.7	140158.7	18606.7	60617.7	23500.72

Table 2 Model predictions vs actual data

Model Performance Among the evaluated machine learning models, the Random Forest Regressor (RF) emerges as the top performer. With an exceptionally high R-squared (R²) score of 0.7470, it excels in explaining approximately 74.70% of the variance in health insurance claim amounts beating Gradient Boosting Regressor (GR) of 0.718 which is 71.8% and XGBoost Regressor (XG): 0.7396 which is 73.96%. This outstanding performance positions the Random Forest Regressor as the most effective model for predicting health insurance claims in our analysis.

4.3 Key EDA Finding

Descriptive Statistics Interpretation: Dataset Size: 16328

Key Insights by Variable: Age, Mean: ~39 years in Range of 18–73 years. Distribution: Right-skewed (median=34, mean=39). On average, clients have claims of around '87372.56'. However, there are also clients whose claims exceed 969674, and these clients are outliers that cause the distribution of the claims column to be right-skewed with a skewness value of 2.774. The distribution of the age column is classified as a normal (symmetrical) distribution with a skewness value close to 0, and the mean and median values are close to each other. It is known that the average age of the clients is around 43 years, with the youngest being 18 years old and the oldest being 73 years old. Looking at the scatter plot, it can be determined that the age column has a positive correlation with the claims column, meaning that as the client's age increases, their claims also tend to increase.

4.4 Feature Importance

- RF: Age most important (45.1%), Acts category (31.7%), BMI (12.6%).

Feature	Importance
Age	0.451323
Category_Acts	0.316585
Bmi	0.126088
Blood_pressure	0.065656
Gender_M	0.009731
Gender_F	0.00918
Regular_ex	0.007851
NCD_yes	0.006591
NCD_no	0.006278
Category_Imaging	0.000552
Category_Consultation	0.00006
Category_Labo	0.000057
Category_Hospitalization	0.000048

- XG: Acts category (63.9%), Age (22.2%), BMI (2.8%).

Feature	Importance
Category_Acts	0.638958
Age	0.22166
NCD_No	0.037473
Bmi	0.028361
Gender_F	0.026226
Regular_ex	0.020183
Blood_pressure	0.018215
Category_Imaging	0.006704
Category_Hospitalization	0.000975
Category_Consultation	0.000646
Category_Labo	0.0006
Gender_M	0
NCD_Yes	0

The analysis of feature importance using Random Forest and XGBoost models reveals key differences in how each model weighs the predictive power of variables. Both models consistently identify Age and Category Acts as the most significant features for predicting the target variable, although their relative importance is inverted between the two algorithms. Random Forest prioritizes Age as the single most important feature (0.451323), followed by Category Acts (0.316585). In contrast, XGBoost assigns a much higher importance to Category Acts (0.638958) and ranks Age as the second most important feature (0.221660). This disparity highlights a crucial difference in the models' underlying mechanisms: Random Forest, an ensemble of decision trees, may be more sensitive to the direct impact of a single strong predictor like Age, whereas XGBoost, a gradient-boosting algorithm, may find more predictive power in a categorical feature like Category Acts, which could be a proxy for a wider range of behaviors or health conditions.

4.5 Interpretation using SHAP

The SHAP value (x-axis) represents how much each feature contributes to the prediction. Positive values (right of 0) increase the predicted output (e.g., higher cost/risk). Negative values (left of 0) decrease the predicted output.

Notable Features:

Demographics: Age and BMI have moderate positive impacts. Category Acts is a medical procedure type that significantly drives costs; its presence or absence can substantially alter predictions. Age and BMI have a more linear, consistent influence (higher values lead to higher costs). Regular ex (exercise) has minimal effect, suggesting lifestyle data might not be as strong a predictor here.

4.6 Discussion

Ensemble tree-based models especially Random Forest outperformed linear regression, confirming their ability to capture non-linear interactions in elderly expenditure patterns. These findings align with international literature (Langenberger, Schulte, & Groene, 2023) and advance methodological approaches in the East African insurance context.

SHAP analysis provided interpretability, showing how age, BMI, and procedure categories jointly influence costs. The indirect role of NCDs suggests preventive interventions and service management could reduce high-cost claims.

Compared with prior East African studies, there is a clear methodological advancement. Works such as (Nyandekwe, Nzayirambaho, & Kakoma, 2020) relied on traditional statistical models (e.g. logistic regression, descriptive analysis), which are valuable for association testing but may not fully capture non-linear, interaction heavy expenditure patterns. By contrast, the ML models applied here not only provided higher predictive accuracy but also offered interpretability through feature importance and SHAP methods. From a policy perspective, the findings confirm the initial hypothesis that the elderly population significantly impacts health insurance expenditures. However, the results also underscore that age is not the sole driver, BMI and service category have equally important roles. This aligns with the Grossman health capital model (Grossman, 1972) which emphasizes that healthcare costs are shaped by a combination of demographic, behavioral, and service related factors.

5. Conclusion & Recommendations

5.1 Conclusion

Machine learning, particularly Random Forest, offers a robust approach to predicting elderly healthcare expenditures and identifying cost drivers in Rwanda's health insurance. In conclusion, this study

contributes to the growing body of evidence that ML based predictive modeling can enhance actuarial decision making in health insurance. It addresses a methodological gap in the Rwandan context by applying nonlinear, interpretable models to elderly expenditure prediction, offering practical implications for risk adjustment, preventive program design, and pricing strategies.

5.2 Limitations

While this study provides important insights into the drivers of health insurance expenditures among the elderly in Rwanda and demonstrates the potential of machine learning models in actuarial analysis, several limitations should be acknowledged:

1. **Data Scope and Representativeness:** The analysis relied solely on claims data from the Military Medical Insurance (MMI), which may not fully reflect expenditure patterns across other insurance schemes or the uninsured elderly population. As such, the findings should be interpreted with caution when generalizing to the broader national context.
2. **Limited Time Frame:** The dataset covers a three-year period, which may not capture longer term trends or seasonal fluctuations in healthcare utilization and costs. Economic shocks, policy changes, or public health emergencies outside this period could influence expenditure patterns differently.
3. **Variable Availability:** Certain potentially influential predictors, such as lifestyle factors, detailed Socio-economic status, and clinical severity of conditions, were not available in the dataset. The absence of these variables may limit the explanatory power of the models.

5.3 Recommendations:

1. Integrate preventive health programs targeting BMI and chronic disease management.
2. Develop clinical guidelines for high-cost service categories (Acts, imaging, hospitalization).
3. Enhance pricing models with ML-derived risk factors beyond age.
4. Build insurer capacity for continuous ML model deployment and updating.

References

1. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, 281–305.
2. Breiman, L. (2001). Random Forests. *Machine Learning*, 1573-0565.
3. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In T. Chen, & C. Guestrin, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Newyork: Association for Computing Machinery.
4. DANGETI, P. (2017). Statistics for Machine Learning: Techniques for exploring supervised, unsupervised, and reinforcement learning models with Python and R. *Packt Publishing*.
5. Gerland, P., Hertog, S., Wheldon, M., Kantorova, V., Gu, D., Gonnella, G., . . . Spoorenberg, T. (2022). *World Population Prospects 2022: Summary of results*. New york: United Nations.
6. Grossman, M. (1972). On the Concept of Health Capital and the Demand for Health. *Journal of Political Economy*, 223-255.
7. Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. *Geoscientific Model Development*, 5481-5487.
8. Jian, G. (2024). R-Squared (R²) – How much variation is explained? *Research Methods in Medicine & Health Sciences*, 105-109.

9. Kajubi, P., Tumushabe, J., & Turyareeba, D. (2021). Aging and household health expenditures in Uganda: An analysis of determinants and policy implications. *International Journal of Health Policy and Management*, 724-733.
10. Kaushik, K., Bhardwaj, A., Dwivedi, A., & Singh, R. (2022). Machine Learning-Based Regression Framework to Predict Health Insurance Premiums. *International journal of environmental research and public health*, 7898.
11. Langenberger, B., Schulte, T., & Groene, O. (2023). The application of machine learning to predict high-cost patients: A performance-comparison of different models using healthcare claims data. *PLoS ONE*, e0279540.
12. Malene, K.-L., Alexander, O. M., Seetha, M., & Jes, S. (2024). Aging populations and expenditures on health. *The Journal of the Economics of Ageing*, 100518.
13. MINALOC. (2021). *National older persons policy*. Kigali: Ministry of Local Government.
14. Mugo, M. G. (2023). The impact of health insurance enrollment on health outcomes in Kenya. *Health Economics Review*, 13-42.
15. Muhanga, M., & Mubyazi, G. M. (2022). Determinants of health care utilization and costs among the elderly in Tanzania: Evidence from household survey data. *BMC Health Services Research*, 1-12.
16. Nyandekwe, M., Nzayirambaho, M., & Kakoma, J. (2020). Universal health insurance in Rwanda: major challenges and solutions for financial sustainability case study of Rwanda community-based health insurance part I. *Pan African Medical Journal*, 37-55.
17. Roger, M., Dominique, H., François, N., & Kabano, I. (2023). A cross-sectional analysis of the impact of health insurance on the use of health care in Rwanda. *Heliyon*, 2405-8440.
18. RSSB. (2024, January 1). *ejo-heza*. Retrieved from RSSB: <https://www.rssb.rw/scheme/ejo-heza>
19. Wang, W., & Lu, Y. (2018). Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model. *IOP Conference Series: Materials Science and Engineering*, 012049.