“Drivers of potential policyholders’ uptake of insurance in Kenya using Random Forest”

AUTHORS	Nelson K. Yego Joseph Nkurunziza Juma Kasozi

ARTICLE INFO	Nelson K. Yego, Joseph Nkurunziza and Juma Kasozi (2023). Drivers of potential policyholders’ uptake of insurance in Kenya using Random Forest. *Insurance Markets and Companies*, *14*(1), 22-34. doi:10.21511/ins.14(1).2023.03
DOI	http://dx.doi.org/10.21511/ins.14(1).2023.03
RELEASED ON	Thursday, 15 June 2023
RECEIVED ON	Wednesday, 22 February 2023
ACCEPTED ON	Friday, 12 May 2023
LICENSE	(cc) BY This work is licensed under a Creative Commons Attribution 4.0 International License
JOURNAL	"Insurance Markets and Companies"
ISSN PRINT	2616-3551
ISSN ONLINE	2522-9591
PUBLISHER	LLC “Consulting Publishing Company “Business Perspectives”
FOUNDER	LLC “Consulting Publishing Company “Business Perspectives”

NUMBER OF REFERENCES	NUMBER OF FIGURES	NUMBER OF TABLES
40	2	5

BUSINESS PERSPECTIVES

LLC "CPC "Business Perspectives"
Hryhorii Skovoroda lane, 10,
Sumy, 40022, Ukraine
www.businessperspectives.org

Nelson K. Yego (Rwanda), **Joseph Nkurunziza** (Rwanda),
Juma Kasozi (Rwanda, Uganda)

DRIVERS OF POTENTIAL POLICYHOLDERS' UPTAKE OF INSURANCE IN KENYA USING RANDOM FOREST

Abstract

The low adoption of insurance by potential policyholders in developing countries like Kenya is a cause for concern for insurers, regulators, and other marketing stakeholders. To effectively design targeted marketing strategies to boost insurance adoption, it is crucial to determine the factors that affect insurance uptake among potential policyholders. In this study, the 2021 FinAccess Survey, which interviewed sampled individuals above 16 years in Kenya and machine learning techniques, including Random Forest, XGBoost, and Logistic Regression, were utilized to uncover the factors driving insurance uptake and the reasons for the low adoption of insurance among potential policyholders. Random Forest was the most robust model of the three classifiers based on Kappa score, recall score, F1 score, precision, and area under the operating characteristic curve (approaching 1). The paper explores eight reasons why people currently do not have insurance policies. The results indicated that affordability was the primary driver of uptake with 68.67% of having expressed a desire to possess insurance but are unable to afford it. The highest level of education being the next most significant factor. Cultural and religious beliefs and mistrust of insurance providers were found to have a minimal impact on uptake. These findings imply that offering affordable insurance products and conducting awareness campaigns are critical to increase insurance adoption.

Received on: 22nd of February, 2023
Accepted on: 12th of May, 2023
Published on: 15th of June, 2023

Nelson K. Yego, Ph.D. Student, African Centre of Excellence in Data Science, University of Rwanda, Rwanda; Department of Mathematics and Computer Science, Moi University, Kenya. (Corresponding author)

Joseph Nkurunziza, Ph.D., Principal, African Centre of Excellence in Data Science, University of Rwanda, Rwanda; School of Economics, University of Rwanda, Rwanda.

Juma Kasozi, African Centre of Excellence in Data Science, University of Rwanda, Rwanda; Ph.D., Professor, Department of Mathematics, Makerere University, Uganda.

Keywords insurance uptake, machine learning, FinAccess data, insured, optimal classifier

JEL Classification G22, G52, C38, D14

INTRODUCTION

Insurance is essential for risk management and providing financial protection (Rumson & Hallett, 2019). However, low insurance penetration and fraudulent claims, particularly in developing countries and other resource constrained environments such as Kenya, remains a concern (Salmi & Atif, 2022; Tessema et al., 2021). Despite a rise in the number of insurance providers and agencies, the percentage of the population that has insurance coverage has remained stagnant at 3.01%. It is important to find a model that could robustly predict the drivers of insurance uptake among potential policyholders using machine learning techniques. The lack of insurance coverage hinders the growth and development of the country, both economically and socially (Mutembei, 2022; Mwongela, 2022). Therefore, identifying the reasons for low uptake and as a result low penetration is imperative for insurers, regulators, and other marketing stakeholders to design effective strategies to promote insurance uptake.

Conflict of interest statement:
Author(s) reported no conflict of interest

http://dx.doi.org/10.21511/ins.14(1).2023.03

1. LITERATURE REVIEW

The review starts off by discussing the factors that affect insurance, then it examines how random forests are used in insurance, and it ends by stating the purpose of this study.

1.1. Drivers of insurance

The factors that drive potential policyholders to purchase insurance are varied and can be analyzed from both the demand and supply perspectives. A demand-side analysis examines factors from the policyholder's viewpoint, such as economic, social, cultural, and regulatory factors. One of the postulations is that the age demographic of a population affects the insurance industry. On the other hand, a supply-side analysis considers the insurer's perspective and identifies factors such as affordability of premiums, Regulations on prices, claims processing procedures, supply paths, and products regulations as having a significant impact on insurance adoption (Dragotă et al., 2022; Mutembei, 2022; Mwongela, 2022; Sibiko & Qaim, 2020). Under supply side factors such as earnings and profitability, reinsurance and actuarial issues, capital adequacy, liquidity, asset quality, and management soundness impact financial soundness of insurers and indirectly supply of the products (Salameh, 2022). Cultural beliefs and superstitions have been shown to affect the uptake of life insurance in certain environments (Liu et al., 2021).

The reasons for low insurance uptake among potential policyholders are complex and multi-faceted, and are influenced by a plethora of factors, including economic, social, cultural, and regulatory dimensions. Understanding the drivers of potential policyholders' likelihood of insurance uptake is essential to promote insurance uptake. Such an analysis could be conducted from either the demand-side or supply-side. In a supply-side analysis, factors such as price regulations, the process of settling insurance claims, the distribution of insurance products, and regulations concerning insurance products have all been identified as being highly influential in determining the chance of insurance being adopted (Ankrah et al., 2021; Mwongela, 2022).

However, a demand-side analysis would provide additional information necessary for data-based decision-making, since it helps to identify the insurer's brand-owned touchpoints. Identifying the insurer's brand-owned touchpoints is essential, since these are the only contact points with potential policyholders and policyholders that the insurers can directly influence. Additionally, analyzing the monetary impact of these touchpoints is also vital to the insurers (Kumar et al., 2023; Zimmermann & Auinger, 2022).

Tessema et al. (2021) emphasize the importance of understanding the risks, potential perils, and the role of insurance in mitigating them in the adoption process. They stress the importance of understanding how the product is being received and adopted by potential customers. However, when they implemented a video intervention, the outcome varied according to the family's head's gender. The intervention increased the use of index insurance for households headed by men by 2-3%, but for households headed by women, it resulted in a 6% decrease in uptake of the insurance.

A phenomenon known as "charity hazard" occurs when households in risk-prone locations decide not to get insurance because they anticipate receiving assistance in the event of a disaster. Coastal families with positive expectations of being eligible for disaster help were 25-42% less likely to have flood insurance, according to research combining household-level survey information with instrumental parameters to investigate flood insurance uptake. This indicates that the expectation of receiving aid may be an important aspect of underinsurance of these communities (Landry et al., 2021).

To gain a greater understanding how government and insurance policies along with management might increase insurance consumption in Zambia, thematic analysis (TA) was used in the research. Financial literacy, excellent service, and regulation of the insurance sector were found to be the three key issues that were essential for encouraging insurance consumption. The study showed that resolving these issues will increase the uptake of insurance. Additionally, the study suggested that simplifying insurance messages for better understanding, providing incentives for insurers to operate in rural areas, and subsidizing certain insurance products would also lead to increased insurance consumption (Haamukwanza, 2021). Other studies have shown that in the insurance sector, customer retention is influenced by reputation, performance, and affect. Additionally, it has been postulated that customer inertia plays a crucial role in moderating the negative effect on health insurance policy customer retention (Iacobucci et al., 2019).

Insurance, being a service industry, requires a strong emphasis on delivering high-quality services, increased recognition of the potential benefits for both the company and the customer, and the integration of advanced technology. Technology, in particular, has had a significant impact on shaping marketing and promotional strategies in the insurance industry (Yadav & Pavlou, 2020). Marketing itself is being disrupted due to the abundance of data and the increasing use of marketing analytics. Fortunately, the disruption could be for the better if correct tools and strategies are employed. Insurers, as well as other businesses, should therefore increase their ability to utilize marketing analytics and metrics as an effective means to gain market insights, monitor and enhance performance. The COVID-19 pandemic is said to have stimulated the digitalization of the insurance industry in Ukraine and other countries (Polinkevych et al., 2022). There is a need to utilize internet marketing and various tools that have come with big data revolution remain competitive (Iacobucci et al., 2019; Prymostka, 2018).

The literature suggests that the drivers of insurance uptake can be analyzed from both the demand-side and supply-side perspectives. Understanding the risks and potential perils, as well as the role of insurance in mitigating them, is crucial in the adoption process. However, the effectiveness of interventions, such as video interventions, may vary depending on factors such as the gender of the leader of the family. Additionally, the "charity hazard" phenomena draw attention to the possibility of detrimental effects of disaster relief expectations on insurance uptake in hazard-prone populations. Additionally, research indicates that addressing financial literacy, service quality, and insurance industry regulation may have a favorable effect on the uptake of insurance, as well as providing incentives for insurers to operate in rural areas and subsidizing certain insurance products.

1.2. Use of Random Forest in insurance

The use of machine learning in this study is based on current and past research that has shown it to be the most robust method for analysis and prediction in similar data (Blier-Wong et al., 2021). Random Forest as a model is part of ensemble tree-based learners, which have shown better performance compared to standalone machine learning models as SVM and other classification methods even when working with imbalanced data (Hanafy & Ming, 2021; Kipkogei et al., 2021; Lin et al., 2017).

The Random Forest technique has been used to develop models that forecast the effectiveness of marketing plans intended to reduce client churn. The type of policy portfolio databases that can be used for a similar function is likewise covered by the proposed model. The study looked at the issue of client churn in the insurance industry. The study recommended locating target customers who are likely to respond favorably to focused marketing and retention efforts rather than concentrating on those with a high risk of departing (Guelman et al., 2012).

Random Forest was used by Shehadeh et al. (2016) to examine data from 130,000 life insurance applications and discovered that stratified sampling was crucial in order to effectively utilize the algorithm. Random Forest was handy for the nature of data because the data was highly imbalanced, with claims making up less than 5% of the total dataset.

According to Guo et al. (2019), Random Forest is the most efficient model for creating a recommender algorithm to suggest insurance products to potential policyholders, when compared to other algorithms such as ID3, C4.5, Naive-Bayes and K-Nearest Neighbors (KNN). The study evaluated the performance of each algorithm using prediction error, which measures the difference between the predicted and actual values. The results of the study indicated that the prediction error of Random Forest was lower than ID3, C4.5, Naive-Bayes and KNN. This suggests that Random Forest is better able to accurately predict the insurance products that a customer may be interested in, when compared to the other algorithms evalu-

http://dx.doi.org/10.21511/ins.14(1).2023.03

ated in the study. The results of the study support methods such as K-Nearest Neighbors (KNN), ID3, C4.5, Naive-Bayes, and others. By evaluating the variation between the expected and actual values, errors in prediction was used in the study to assess individual algorithm's performance. The study's findings showed that Random Forest had a smaller prediction error than ID3, C4.5, Naive-Bayes, and KNN. This shows that, when compared to the other algorithms considered in the study, Random Forest is better able to forecast with accuracy the insurance products that a consumer may be interested in. The study's findings confirm the use of Random Forest as a suitable algorithm for creating a recommender system for insurance products.

Hanafy and Ming (2021) examined the adoption of machine learning in the field of automotive insurance and also explored its potential applications for handling large amounts of data. To forecast the occurrence of claims, the study used a variety of machine learning techniques, including logistic regression, Extreme Gradient Boosting (XGBoost), Random Forest, decision trees, Naive Bayes, and K-Nearest Neighbors (KNN). These models' performances were assessed and contrasted, and the findings revealed that Random Forest performed better than the alternative techniques in terms of accuracy, kappa, and AUC values.

It has been postulated that for insurers to be successful and competitive in the market that is currently undergoing a big data revolution, they must utilize the increasingly large amounts of data in their decision-making. One of such applications is in targeted marketing to customize insurance policies (Porrini, 2017). This paper builds on this concept by using recent data to analyze sociodemographic data to find optimal model for analyzing the drivers of policyholders' uptake.

Salmi and Atif (2022) proposed a data mining methodology to identify false claims by addressing class imbalance and experimenting with two alternative feature subsets. It used two sampling techniques: SMOTE and ROSE. The findings demonstrated that the models that were created utilizing the second feature selection performed marginally better, with a higher percentage of false claims accurately detected. Random Forest outperformed logistic regression, according to the study.

In conclusion, studies discussed in the literature have shown that Random Forest could be an effective algorithm for a variety of tasks related to the insurance industry, such as predicting customer churn, identifying target customers for targeted marketing and retention efforts, creating a recommender system for insurance products, and detecting fraudulent claims. The studies have also shown that in terms of accuracy and prediction error, Random Forest surpasses other algorithms such as logistic regression, ID3, C4.5, Naive-Bayes, and K-Nearest Neighbors. Additionally, the studies have highlighted the importance of addressing class imbalance and utilizing effective sampling methods, such as SMOTE and ROSE, when working with imbalanced datasets. Overall, the literature supports the use of Random Forest as a suitable algorithm for various tasks in the insurance industry.

This study aims to uncover the drivers of the insurance uptake and reasons behind the low uptake of insurance among potential policyholders in Kenya from a demand-side perspective using the optimal classifier.

2. METHODS

The study determines the factors driving insurance uptake using data from the 2021 FinAccess Survey. The study utilized the most recent FinAccess data, which is part of a series of national surveys conducted to evaluate the access, usage, and impact of financial inclusion. The initial step involved using frequency analysis to identify the reasons why people currently do not have insurance policies. Subsequently, a machine learning model was trained and tested to extract the most important variables affecting insurance uptake. The feature importance was then extracted from the model to determine the variables that have the greatest influence on insurance uptake.

The 2021 FinAccess Survey is the sixth in a series that began in 2006. The survey includes measures of consumer protection to assess not only access to and use of finance but also the impact of financial inclusion on people's financial wellbeing. The study used a cross-sectional design at the household level and targeted people aged 16 and up living in traditional Kenyan families. The set of respondents was obtained by utilizing the Kenya Household Master Sampling Frame, and this originated from the 2019 Kenya Population and Housing Census to provide national, rural/urban, and regional estimations. The survey's minimal number of respondents had been determined for each domain, resulting in 1,700 enumeration regions or clustered and 30,600 households. A representative sample at the national and county levels was produced using the data after it had been processed and inability to respond adjustments were made. 25,724 members of the sample were qualified for interviews at the time of data collection, and 22,024 of those surveys had a positive outcome, resulting in a general response rate of 85.6%, with rural areas responding at an 88.6% rate and urban areas responding at an 80% rate (Kenya National Bureau of Statistics, 2021).

Feature selection techniques are helpful in improving the correctness of classification algorithms by lowering the dimensionality of the datasets (Li et al., 2017; Rawat et al., 2021). The original dataset consisted of over a thousand variables, however, only 29 socio-demographic features that were most relevant to the respondents and the goals of the study were selected. The features with less than 30% missing observations were later removed. The unselected features were either not socio-demographic, had more than 30% missing observations, or were unnecessary as they were already explained by the chosen features.

One of the goals of the study was to identify the most accurate machine learning classifier for predicting insurance uptake by potential policyholders. Three traditional classifiers evaluated were the Random Forest, XGBoost, and Logistic Regression, and the best classifier was selected for the final analysis. Random Forest and XGBoost are ensemble models that use trees, while Logistic Regression is a popular model for binary prediction results (Ampomah et al., 2020; Basak et al., 2019).

The Random Forest classifier is a machine learning algorithm that is commonly used for classification and regression applications. The Random Forest algorithm creates many decision trees and combines their outputs to make a final prediction. The decision trees are created by selecting random subsets of the data and features. The anticipated outcomes of all the choices are taken into account to provide the most accurate prediction. Random Forest ensembles trees through bagging. Bagging reduces the high variance of decision tree classifiers by creating new datasets from the original dataset. In Random Forest, each new dataset contains a number of observations say "n", which are drawn at random from the original dataset with replacement. The predictions from each tree are then averaged together after each fresh dataset is used to generate a classification tree. This results in an ensemble estimate of the classification function.

$$f_{av}\left(X\right)=\frac{1}{N}\sum\nolimits_{n=1}^{N}f_n\left(X\right), \tag{1}$$

where f_n is the A classification function that fits the classification tree to the nth new dataset (Diana et al., 2019).

In the current study, the Random Forest was used to classify potential policyholders based on their likelihood of taking up insurance. Various socio-demographic variables, such as age, income, and education level, were employed to train the model, which was subsequently utilized for predicting which potential policyholders were more likely to take up insurance. The Random Forest algorithm's the capacity to manage a variety of features, missing values, and its robustness to noise in the data make it an ideal algorithm for this research, as it can handle high-dimensional data and missing values which are prevalent in the FinAccess datasets (Hou et al., 2020; Ren et al., 2023).

The first step was to perform a frequency analysis to determine the reasons why individuals did not have insurance policies. Frequency tables were created based on specific answers to questions regarding why the survey respondents did not have insurance at the time of the survey.

The data was then divided into three parts in a 70:15:15 ratio for training, testing, and validation. The split ratio was based on other studies of similar nature (Ding et al., 2020; Kipkogei et al., 2021; Wu et al., 2022). To make reproducible results as advised in (Kassambara, 2018), a seed of 2 was set, and a The 5-fold cross-validation approach was employed. In insurance-related investigations, k-fold cross validation has been successful (Quan et al., 2023). The training was first performed on the imbalanced data, then SMOTE and up sampling methods were applied for sampling. The k-fold cross-validation was conducted using the validation set served to test the models' performance, and the test set was employed to evaluate the models' performance. The reported metrics refer to the results from the test set.

3. RESULTS

The results from frequency analysis are first presented, followed by the performance of the three models under various data sampling techniques, and finally, the results of feature importance are presented.

3.1. Frequency analysis results

Table 1 shows that 68.67% of respondents (potential policyholders to insurers) in the study reported that they would like to have insurance but cannot afford it, while 0.21% believed that buying health or life insurance brings bad luck. Additionally, 10.93% said that they do not know where to obtain insurance, and 0.86% believed that insurance companies are dishonest. A smaller percentage of respondents, 0.48%, reported that they believe insurance agents are dishonest. A small number of respondents, 0.86%, stated that they do not need insurance, and 0.41% reported that they save for emergencies instead of purchasing insurance. Lastly, 0.47% cited religious or cultural reasons for not having insurance.

3.2. Model metrics

3.2.1. Model metrics on imbalanced data

Table 2 presents a performance comparison of three models (Logistic Regression, Random Forest, and XGBoost) on an imbalanced data set using four evaluation metrics: Kappa, Recall, F1 score, and Accuracy. The higher the Kappa score, the more effective the model's performance. The Random Forest model performs better than both of the two models with the highest Kappa score (0.158534), the highest recall score (0.548966), the highest F1 score (0.569272), and the highest accuracy score (0.926755). In regard to Kappa, the Random Forest model provides the best performance as a whole, Recall, F1, and Accuracy metrics on the imbalanced data set.

Table 1. Reasons why people currently do not have an insurance policy

Question	Valid response	Frequency	Percentage
You would like to have insurance but cannot afford it	No	6,685,011	31.33
	Yes	14,655,637	68.67
Attempting to purchase health or life insurance for yourself or your family might be unlucky.	No	21,296,279	99.79
	Yes	44,369	0.21
You do not know where to get it from	No	19,008,790	89.07
	Yes	2,331,858	10.93
Insurance companies are dishonest	No	21,157,450	99.14
	Yes	183,198	0.86
Insurance agents are dishonest	No	21,237,917	99.52
	Yes	102,731	0.48
You do not need insurance	No	21,157,815	99.14
	Yes	182,833	0.86
Because you save for emergencies, you do not have insurance.	No	21,253,551	99.59
	Yes	87,097	0.41
Religious reasons /cultural reasons	No	21,241,306	99.53
	Yes	99,342	0.47
Total for each case	–	21,340,648	100.00

Table 2. Model metrics on imbalanced data

N	Model	Kappa score	Recall score	F1 score	Accuracy
0	Logistic Regression	0.013273	0.503641	0.487976	0.923729
1	Random Forest	0.158534	0.548966	0.569272	0.926755
2	XGBoost	0.164047	0.553954	0.574593	0.922518

Table 3. Model metrics on SMOTE balanced data

N	Model	Kappa score	Recall score	F1 score	Accuracy
0	Logistic Regression	0.463776	0.731781	0.731796	0.732053
1	Random Forest	0.847329	0.923784	0.923655	0.923655
2	XGBoost	0.869929	0.934906	0.934963	0.934983

3.2.2. Model metrics on SMOTE balanced data

Table 3 presents a comparison of the effectiveness of all the three models, including Logistic Regression, Random Forest, and XGBoost on an SMOTE balanced data set using four evaluation metrics: Kappa, Recall, F1 score, and Accuracy. The XGBoost model has the highest Kappa score (0.869929) among the three models, which is then followed by the Random Forest (0.847329). The XGBoost has the highest recall rating as well (0.934906), and then the Random Forest (0.923784). The XGBoost model has the best balance between precision and recall, as measured by the F1 score (0.934963), followed by the Random Forest (0.923655). In terms of accuracy, the XGBoost model also has the highest score (0.934983), followed by the Random Forest (0.923655). Overall, the XGBoost model performs better than the Logistic Regression and Random Forest models in terms of Kappa, Recall, F1, and Accuracy metrics on this SMOTE balanced data set.

3.2.3. Model metrics on randomly oversampled data

Table 4 indicates that in the oversampled data set, the Random Forest model has the highest Kappa score (0.992121) followed by the XGBoost model (0.820974). The Random Forest model also has the highest recall score (0.996128), followed by the XGBoost model (0.911058). Additionally, the Random Forest model has the highest F1 score (0.99606) and accuracy score (0.996061), followed by the XGBoost model (0.910341 and 0.910389, respectively). As a result, the Random Forest model outperforms the Logistic Regression and XGBoost models in terms of Kappa, Recall, F1, and Accuracy metrics on this oversampled data set.

Table 4. Model metrics on oversampled data

N	Model	Kappa score	Recall score	F1 score	Accuracy
0	Logistic Regression	0.338967	0.669463	0.66948	0.669621
1	Random Forest	0.992121	0.996128	0.99606	0.996061
2	XGBoost	0.820974	0.911058	0.910341	0.910389

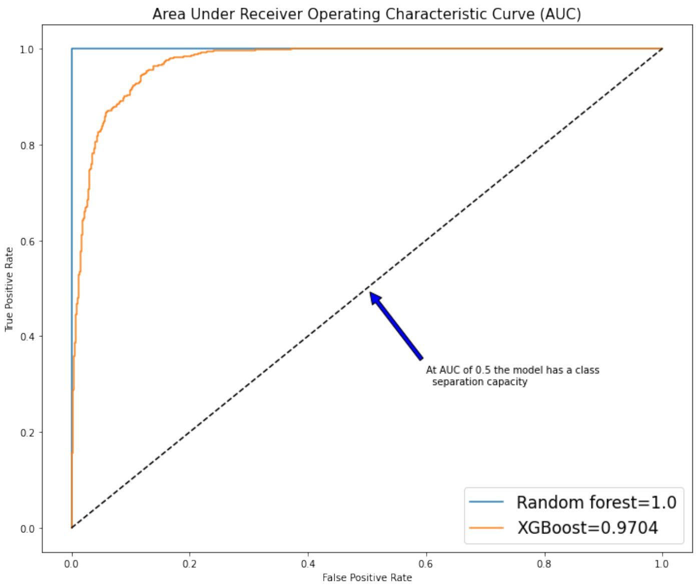

Figure 1. Areas under the ROC curve (AUC)

http://dx.doi.org/10.21511/ins.14(1).2023.03

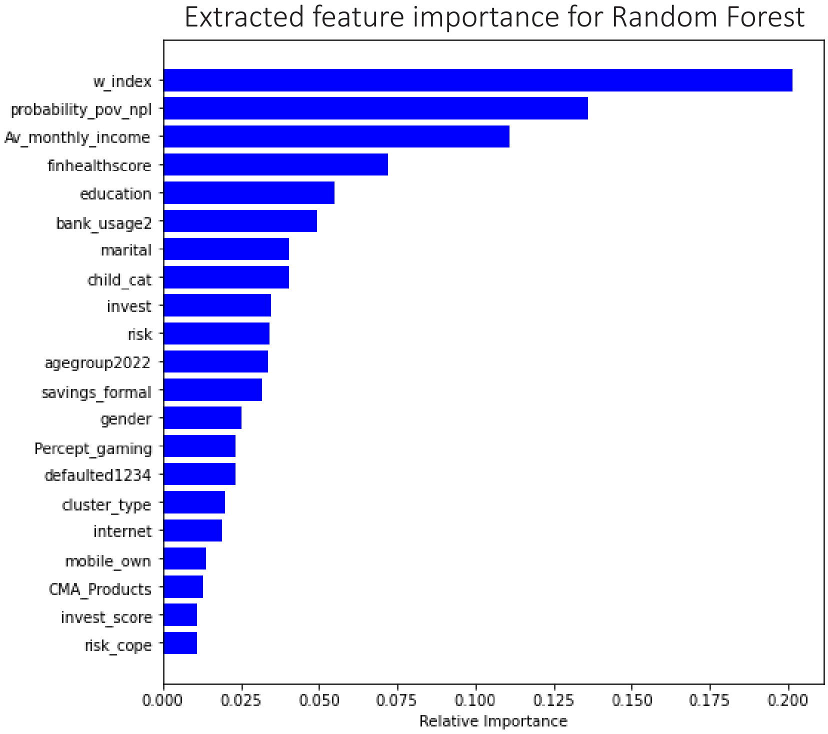

Figure 2. Feature importance

Given the favorable performance of both XGBoost and Random Forest, the AUC metric was introduced to determine the optimal model. The Random Forest model had an AUC value of 1.0, indicating a perfect classifier, for prediction of insurance uptake among potential policyholders, meaning it can distinguish between positive and negative cases with 100% accuracy. Meanwhile, XGBoost has an AUC value of 0.9704, indicating a good performance but not a perfect one.

Figure 2 displays the relative importance of the variables in predicting the uptake of insurance by potential policyholders in Kenya, as determined by the Random Forest model. Wealth quintile level and poverty vulnerability were found to be the most important factors, with relative importance of 0.2014 and 0.1361, respectively. Other factors such as average monthly income, financial health score, and highest level of education attained were also found to be significant, but to a lesser extent. Other variables such as marital status, number of children in the household, and gender of the respondent were found to have low relative importance in predicting the uptake of insurance. The results of the model suggest that the financial status and vulnerability of potential policyholders play a significant role in determining their uptake of insurance in Kenya.

4. DISCUSSION

The performance of three models (Logistic Regression, Random Forest, and XGBoost) is compared on three data sets (imbalanced, SMOTE balanced, and random oversampled) using four evaluation metrics (Kappa, Recall, F1 score, and Accuracy). The best performance is found to be different in each data set. On the imbalanced data set, the Random Forest model outperforms the others. On the SMOTE balanced data set, the XGBoost model performs better. On the oversampled data set, the Random Forest model performs the best. An additional metric, the area under The ROC Curve (AUC) serves to choose the best model between XGBoost and Random Forest. Random Forest is the most reliable model for forecasting the factors that influence potential policyholders' insurance purchase decisions. This corroborates with Hanafy and Ming (2021), Lin et al. (2017), Guo et al. (2019) and Salmi and Atif (2022), who found Random Forest to outperform other models. Therefore, a Random Forest model could be utilized in targeted marketing to predict insurance uptake.

The four most important factors that drive insurance uptake, according to the variable importance, are wealth index, vulnerability to poverty, average monthly income, and financial health score. These four features all point to a single key factor: affordability. The results of the analysis into why individuals currently do not have insurance policies reveal that affordability is the main obstacle, with 68.67% of respondents reporting that they would like to have insurance but cannot afford it. While other factors may play a role in insurance uptake, affordability appears to be the most critical. It has been previously observed that in the insurance industry, potential policyholders seek insurance products and services that are both cost-effective and of high quality (Ali & Tausif, 2018; Kumar et al., 2023). This supports the findings of Sibiko and Qaim (2020), who posited that providing premium subsidies had a direct positive effect on uptake, even if knowledge of the product did not necessarily lead to increased uptake of index-based livestock insurance. However, in the current findings, the results depicted affordability as the most important factor that the potential policyholder considers in product selection. This highlights the importance of designing affordable insurance products.

The highest level of education attained by the respondent is the next most important factor in determining affordability, according to the Random Forest model. This confirms previous findings that insurance uptake increases with education due to increased awareness of insurance and financial products (Mutembei, 2022). The analysis of individuals without insurance policies found that 10.93% do not know where to obtain insurance, highlighting a need for increased awareness. Programs to increase awareness of insurance and its importance to individuals should be intensified.

Gender, internet usage, and mobile phone ownership do not display as much importance as wealth index, vulnerability to poverty, average monthly income, financial health score, and level of education. This suggests a reduction in disparity between gender and the high portion of the population owning phones, and such ownership does not imply much about the demand for insurance.

A small percentage of respondents, 0.21% of respondents in the study, believed that buying health or life insurance brings bad luck, and 0.47% cited religious or cultural reasons for not having insurance. This contradicts the results of Liu et al. (2021), which found superstition to be a significant factor in the uptake of life insurance. This suggests that cultural superstition may not be a hindrance to insurance uptake in Kenya and highlights the need to understand cultural and societal factors that influence insurance uptake in different settings. This information is important for developing effective strategies to promote insurance coverage.

Previously, it was reported that insurance companies and their agents were not honest (Barnes et al., 2010), but only 0.86% of respondents believed that insurance companies are dishonest and 0.48% believed that insurance agents are dishonest. This low perception of dishonesty among insurers and agents may not contribute to the low uptake of insurance. It is possible that the regulatory programs have reduced the instances of dishonesty among insurers and agents to a significant extent.

The study's findings on the drivers of low insurance uptake among potential policyholders in Kenya will be useful for insurers, regulators, and other stakeholders to design effective policies and strategies to promote insurance uptake. The research may have broader implications for other developing countries and resource-constrained environments facing similar challenges in increasing insurance uptake. The study's specific insights into the drivers of low insurance uptake in Kenya can inform policies and strategies in other countries with similar characteristics, allowing policymakers to design targeted marketing interventions that will increase insurance uptake and improve financial security for citizens.

CONCLUSION

The study sought to discover the factors influencing insurance purchases among prospective policyholders and find a model that could robustly predict those factors using machine learning techniques. The study found that Random Forest presented best results amongst the three models that ware tested

http://dx.doi.org/10.21511/ins.14(1).2023.03

and was identified as the most effective model for predicting the factors influencing potential policyholders' uptake of insurance based on the Kappa score, Recall score, F1 score, Accuracy, and Area Under the ROC Curve (AUC) metric. As a result, the Random Forest model would be an ideal choice for developing an algorithm for targeted marketing aimed at increasing insurance uptake among potential policyholders. Based on the results, cost effectiveness was discovered to be the primary driver of insurance uptake in Kenya, with 68.67% of respondents indicating that they cannot afford insurance but would like to have it. The next most significant factor was the respondents' level of education, which was associated with increased awareness of insurance and financial products. While a small percentage of respondents cited cultural and religious reasons or superstitions as barriers to uptake, the data suggested that cultural and societal factors may not be significant barriers in Kenya. Additionally, the low perception of dishonesty among insurers and agents implies that this is not a significant factor in the low uptake of insurance. These findings emphasize the need to design affordable insurance products, increase awareness of insurance and its importance through targeted programs, and understand cultural and societal factors that influence insurance uptake in different settings to promote insurance coverage effectively.

AUTHOR CONTRIBUTIONS

Conceptualization: Nelson K. Yego, Juma Kasozi, Joseph Nkurunziza.
Data curation: Nelson K. Yego.
Formal analysis: Nelson K. Yego.
Funding acquisition: Juma Kasozi, Joseph Nkurunziza.
Investigation: Nelson K. Yego, Juma Kasozi, Joseph Nkurunziza.
Methodology: Nelson K. Yego, Juma Kasozi, Joseph Nkurunziza.
Project administration: Juma Kasozi, Joseph Nkurunziza.
Resources: Nelson K. Yego, Juma Kasozi, Joseph Nkurunziza.
Software: Nelson K. Yego.
Supervision: Juma Kasozi, Joseph Nkurunziza.
Validation: Nelson K. Yego, Juma Kasozi, Joseph Nkurunziza.
Visualization: Nelson K. Yego.
Writing – original draft: Nelson K. Yego.
Writing – review & editing: Nelson K. Yego, Juma Kasozi, Joseph Nkurunziza.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

1. Ali, A., & Tausif, M. R. (2018). Service quality, customers' satisfaction, and profitability: an empirical study of Saudi Arabian insurance sector. *Investment Management and Financial Innovations, 15*(2), 232-247. https://doi.org/10.21511/imfi.15(2).2018.21
2. Ampomah, E. K., Qin, Z., & Nyame, G. (2020). Evaluation of Tree-Based Ensemble Machine Learning Models in Predicting Stock Price Direction of Movement. *Information, 11*(6), 332. https://doi.org/10.3390/info11060332
3. Ankrah, D. A., Kwapong, N. A., Eghan, D., Adarkwah, F., & Boateng-Gyambiby, D. (2021). Agricultural insurance access and acceptability: examining the case of smallholder farmers in Ghana. *Agriculture & Food Security, 10*(1), 19. https://doi.org/10.1186/s40066-021-00292-y
4. Barnes, J., O'Hanlon, B., Feeley III, F., McKeon, K., Gitonga, N., & Decker, C. (2010). *Private health sector assessment in Kenya* (Working Paper No. 193). World Bank Publications. Retrieved from http://documents.worldbank.org/curated/en/434701468048274776/Private-health-sector-assessment-in-Kenya

5. Basak, S., Kar, S., Saha, S., Khaidem, L., & Dey, S. R. (2019). Predicting the direction of stock market prices using tree-based classifiers. *The North American Journal of Economics and Finance, 47,* 552-567. https://doi.org/10.1016/j.najef.2018.06.013

6. Blier-Wong, C., Cossette, H., Lamontagne, L., & Marceau, E. (2021). Machine Learning in P & C Insurance: A Review for Pricing and Reserving. *Risks, 9*(1), 4. https://doi.org/10.3390/risks9010004

7. Diana, A., Griffin, J. E., Oberoi, J. S., & Yao, J. (2019). *Machine-Learning Methods for Insurance Applications-A Survey*. Society of Actuaries. Retrieved from https://www.soa.org/493479/globalassets/assets/files/resources/research-report/2019/machine-learning-methods.pdf

8. Ding, K., Lev, B., Peng, X., Sun, T., & Vasarhelyi, M. A. (2020). Machine learning improves accounting estimates: evidence from insurance payments. *Review of Accounting Studies, 25*(3), 1098-1134. https://doi.org/10.1007/s11142-020-09546-9

9. Dragotă, I.-M., Cepoi, C. O., & Ştefan, L. (2022). Threshold effect for the life insurance industry: evidence from OECD countries. *The Geneva Papers on Risk and Insurance - Issues and Practice.* https://doi.org/10.1057/s41288-022-00272-8

10. Guelman, L., Guillén, M., & Pérez-Marín, A. M. (2012). Random Forests for Uplift Modeling: An Insurance Customer Retention Case. *Lecture Notes in Business Information Processing, 115,* 123-133. http://dx.doi.org/10.1007/978-3-642-30433-0_13

11. Guo, Y., Zhou, Y., Hu, X., & Cheng, W. (2019). Research on Recommendation of Insurance Products Based on Random Forest. *2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)* (pp. 308-311). Taiyuan, China. https://doi.org/10.1109/MLBDBI48998.2019.00069

12. Haamukwanza, C. L. (2021). To insure or not to insure – the role that government and insurance practice should play: a thematic comparison of the urban poor and the workers in the pensions and insurance industry. *SN Business & Economics, 1*(9), 114. https://doi.org/10.1007/s43546-021-00121-1

13. Hanafy, M., & Ming, R. (2021). Machine Learning Approaches for Auto Insurance Big Data. *Risks, 9*(2), 42. https://doi.org/10.3390/risks9020042

14. Hou, Q., Liu, Y., Liu, J., & Sun, S. (2020). Epilepsy Detection Using Random Forest Classification Based on Locally Linear Embedding Algorithm. *2020 5th International Conference on Control, Robotics and Cybernetics (CRC)* (pp. 202-206). https://doi.org/10.1109/CRC51253.2020.9253455

15. Iacobucci, D., Petrescu, M., Krishen, A., & Bendixen, M. (2019). The state of marketing analytics in research and practice. *Journal of Marketing Analytics, 7*(3), 152-181. https://doi.org/10.1057/s41270-019-00059-2

16. Kassambara, A. (2018). *Machine learning essentials: Practical guide in R.* CreateSpace Independent Publishing Platform.

17. Kenya National Bureau of Statistics. (2021). *FinAccess Household Survey 2021*. Retrieved from https://finaccess.knbs.or.ke/reports-and-datasets

18. Kipkogei, F., Kabano, I. H., Murorunkwere, B. F., & Joseph, N. (2021). Business success prediction in Rwanda: a comparison of tree-based models and logistic regression classifiers. *SN Business & Economics, 1*(8), 101. https://doi.org/10.1007/s43546-021-00104-2

19. Kumar, A. N., Girish, S., & Suresha, B. (2023). Switching intention and switching behavior of adults in the non-life insurance sector: Mediating role of brand love. *Insurance Markets and Companies, 14*(1), 1-7. https://doi.org/10.21511/ins.14(1).2023.01

20. Landry, C. E., Turner, D., & Petrolia, D. (2021). Flood Insurance Market Penetration and Expectations of Disaster Assistance. *Environmental and Resource Economics, 79*(2), 357-386. https://doi.org/10.1007/s10640-021-00565-x

21. Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM Computing Surveys (CSUR), 50*(6), 1-45. https://doi.org/10.1145/3136625

22. Lin, W., Wu, Z., Lin, L., Wen, A., & Li, J. (2017). An ensemble random forest algorithm for insurance big data analysis. *IEEE Access, 5,* 16568-16575. https://doi.org/10.1109/ACCESS.2017.2738069

23. Liu, Y., Zhang, Y., Chen, X., & Yang, Y. (2021). Superstition and farmers' life insurance spending. *Economics Letters, 206,* 109975. https://doi.org/10.1016/J.ECONLET.2021.109975

24. Mutembei, J. M. (2022). Impact of Employees Capability Affecting the Growth of Life Insurance Business. A Critical Literature Review. *Journal of Actuarial Research, 1*(1), 1-12. https://doi.org/10.47941/jar.1041

25. Mwongela, J. N. (2022). *The Influence of Regulatory Framework on Insurance Penetration in Kenya. A Case Study of the Registered Insurance Companies in Nairobi County* (Master's Thesis). Kenya Methodist University. Retrieved from http://repository.kemu.ac.ke/handle/123456789/1325

26. Polinkevych, O., Glonti, V., Baranova, V., Levchenko, V., & Yermoshenko, A. (2022). Change of business models of Ukrainian insurance companies in the conditions of COVID-19. *Insurance Markets and Companies, 12*(1), 83-98. https://doi.org/10.21511/ins.12(1).2021.08

27. Porrini, D. (2017). Regulating Big Data effects in the European insurance market. *Insurance Markets and Companies, 8*(1), 6-15. http://dx.doi.org/10.21511/ins.08(1).2017.01

28. Prymostka, O. (2018). Life insurance companies marketing strategy in the digital world. *Insurance Markets and Companies,* *9*(1), 70-78. http://dx.doi.org/10.21511/ins.09(1).2018.06

29. Quan, Z., Wang, Z., Gan, G., & Valdez, E. A. (2023). On hybrid tree-based methods for short-term insurance claims. *Probability in the Engineering and Informational Sciences, 37*(2), 597-620. https://doi.org/10.1017/S0269964823000074

30. Rawat, S., Rawat, A., Kumar, D., & Sabitha, A. S. (2021). Application of machine learning and data visualization techniques for decision support in the insurance sector. *International Journal of Information Management Data Insights, 1*(2), 100012. https://doi.org/10.1016/j.jjimei.2021.100012

31. Ren, L., Seklouli, A. S., Zhang, H., Wang, T., & Bouras, A. (2023). An adaptive Laplacian weight random forest imputation for imbalance and mixed-type data. *Information Systems, 111,* 102122. https://doi.org/10.1016/j.is.2022.102122

32. Rumson, A. G., & Hallett, S. H. (2019). Innovations in the use of data facilitating insurance as a resilience mechanism for coastal flood risk. *Science of The Total Environment, 661,* 598-612. https://doi.org/10.1016/j.scitotenv.2019.01.114

33. Salameh, H. (2022). An Evaluation of the financial soundness of insurance firms in the Amman Stock Exchange. *Insurance Markets and Companies*, *13*(1), 11-20. http://dx.doi.org/10.21511/ins.13(1).2022.02

34. Salmi, M., & Atif, D. (2022). Using a Data Mining Approach to Detect Automobile Insurance Fraud. *Proceedings of the 13th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2021)* (pp. 55-66). Springer International Publishing. https://doi.org/10.1007/978-3-030-96302-6_5

35. Shehadeh, M., Kokes, R., & Hu, G. (2016). *Variable Selection Using Parallel Random Forest for Mortality Prediction in Highly Imbalanced Data* (pp. 13-16). Society of Actuaries. Retrieved from https://www.soa.org/essays-monographs/research-2016-predictive-analytics-call-essays.pdf

36. Sibiko, K. W., & Qaim, M. (2020). Weather index insurance, agricultural input use, and crop productivity in Kenya. *Food Security, 12*(1), 151-167. https://doi.org/10.1007/s12571-019-00987-y

37. Tessema, Y. A., Hobbs, A., & Jensen, N. (2021). *The Role of Learning Styles in the Uptake of Index Insurance: Evidence from Kenya* (Master's Theses). University of San Francisco. Retrieved from https://repository.usfca.edu/thes/1374

38. Wu, K., Wu, E., DAndrea, M., Chitale, N., Lim, M., Dabrowski, M., Kantor, K., Rangi, H., Liu, R., Garmhausen, M., Pal, N., Harbron, C., Rizzo, S., Copping, R., & Zou, J. (2022). Machine Learning Prediction of Clinical Trial Operational Efficiency. *The AAPS Journal, 24*(3), 57. https://doi.org/10.1208/s12248-022-00703-3

39. Yadav, M. S., & Pavlou, P. A. (2020). Technology-enabled interactions in digital environments: a conceptual foundation for current and future research. *Journal of the Academy of Marketing Science, 48*(1), 132-136. https://doi.org/10.1007/s11747-019-00712-3

40. Zimmermann, R., & Auinger, A. (2022). Developing a conversion rate optimization framework for digital retailers – case study. *Journal of Marketing Analytics, 11,* 233-243. https://doi.org/10.1057/s41270-022-00161-y

APPENDIX A

Table A1. Feature importance

Feature label	Explanation	Importance
w_index	Wealth quintile index	0.2014
probability_pov_npl	Poverty vulnerability	0.1361
Av_monthly_income	Average Monthly Income (in Kenyan Shillings)	0.1112
Finhealthscore	Financial health score	0.0721
Education	Highest level of education level attained	0.0551
bank_usage2	Having a formal bank account	0.0495
Marital	Marital status	0.0403
child_cat	number of children in Household	0.0402
Invest	Usage of investment	0.0345
Risk	Risk score	0.0343
agegroup2022	Age group of the respondent	0.0336
savings_formal	Savings usage	0.0320
Gender	Gender of the respondent	0.0254
Percept_gaming	Perception towards gaming	0.0233
defaulted1234	Defaulted in any loan payment	0.0231
cluster_type	Place of residence rural or urban	0.0199
Internet	Having access to internet	0.0189
mobile_own	Ownership of a mobile phone	0.0137
CMA_Products	Usage of securities investment products	0.0131
invest_score	Ability to invest in others' livelihoods	0.0112
risk_cope	Able to cope with risk	0.0109