Article

A Comparative Analysis of Machine Learning Models for Prediction of Insurance Uptake in Kenya

Nelson Yego [1,2*], Juma Kasozi [1,3] and Joseph Nkurunziza [1,4]

[1] African Center of Excellence in Data Science, Rwanda; nelsonyego@gmail.com
[2] Department of Mathematics and Computing, Moi University, Kenya; nelsonky@mu.ac.ke
[4] Department of Mathematics, Makere University, Uganda; kasozi@cns.mak.ac.ug
[3] School of Economics, University of Rwanda, Rwanda; nkurunzizaj@gmail.com
* Correspondence: nelsonyego@gmail.com

Abstract: The role of insurance in financial inclusion as well as in economic growth is immense. However, low uptake seems to impede the growth of the sector hence the need for a model that robustly predicts uptake of insurance among potential clients. In this research, we compared the performances of eight (8) machine learning models in predicting the uptake of insurance. The classifiers considered were Logistic Regression, Gaussian Naive Bayes, Support Vector Machines, K Nearest Neighbors, Decision Tree, Random Forest, Gradient Boosting Machines and Extreme Gradient boosting. The data used in the classification was from the 2016 Kenya FinAccess Household Survey. Comparison of performance was done for both upsampled and downsampled data due to data imbalance. For upsampled data, Random Forest classifier showed the highest accuracy and precision compared to other classifiers but for downsampled data, Extreme Gradient boosting was optimal. It is noteworthy that for both upsampled and downsampled data, tree-based classifiers were more robust than others in insurance uptake prediction. However, in spite of hyper-parameter optimization, the Area under Receiver Operating Characteristic curve (AUC) remained highest for Random Forest as compared to other tree-based models. Also, the confusion matrix for Random Forest showed the least false positives, and the highest true positives hence could be construed as the most robust model for predicting the insurance uptake. Finally, the most important feature in predicting uptake was having a bank product hence bancassurance could be said to be a plausible channel of distribution of insurance products.

Keywords: Insurance Uptake, Machine Learning, Upsample, Downsample.

1. Introduction

1.1. Background of the Study

The role of insurance in financial inclusion as well as in sustainable economic growth is immense. Insurance is not only important for its risk pooling and transfer roles but also for capital accumulation for investment. Nevertheless, less discussion had been canvassed about its importance. Much of the discussion has been on the banking sector but less on the insurance sector [14]. Since African Union Agenda 2063 envisions inclusive sustainable development in Africa, insurance could contribute a great deal to achieving this. This is more so due to the tendency of financial risks to increase over time. Therefore, sustainability of the growth could be cushioned by hedging the risks and in this, insurers are better placed for their ability to indemnify their clients in case of a peril operation [7,12].

Despite its role in financial inclusion and sustainable economic growth, low insurance uptake and low penetration have seemed to impede growth of the insurance sector. Insurance uptake has been low in Kenya. This was regardless of various programs developed and implemented by the Association of Kenya Insurers (AKI) to increase the uptake and consequently the penetration. Moreover, life insurance penetration in the Kenyan market had worsened for the third year in a row, dipping to 2.79% in 2015 from 2.94% in 2014. This was notwithstanding the fact that the gross domestic product (GDP) grew to 5.6% in 2015 in comparison to 5.3% in 2014. The low uptake has spanned most lines of insurance. Uptake of insurance has been one of the impediments to growth and expansion of insurance. This is possibly because low uptake directly affects financial performance since the lower the uptake the lower the penetration and consequently the lower the financial performance [1].

The uptake could be investigated from the demand side of insurance in terms of the count of people who take up policies, while penetration could be thought of as premiums as a ratio of gross domestic product. Nevertheless, both ways of looking at insurance growth had shown low statistics. The problem of low insurance uptake could be alleviated by targeted uptake promotion, hence the need for a model that robustly predicts uptake. Finding an optimal and robust way of predicting insurance policy uptake can help determine whether a customer or a potential one will take up insurance. Such a model would be valuable from the insurers point of view, especially in putting in place a distribution strategy for insurance. This will help in targeted marketing and product improvement. [20]. Previous studies related to this paper focused on a particular line of insurance and none made use of the machine learning approach. For instance, the uptake of private health insurance, community based kinds of health insurance, remote sensing insurance, flood insurance and agricultural insurance. Whereas these are vital in ensuring in-depth knowledge of these particular lines, there is a need to have a look at the overall picture of insurance as a whole Lambregts and Schut [12]. This study attempted to fill this gap by having a peep into the insurance uptake prediction from an overall perspective by considering all the lines of insurance. The labels are therefore related to insurance uptake on all the lines: life, non-life and health insurance.

The use of machine learning is motivated not only by its predictive capabilities but also by the documented possibility of yielding new insights. Moreover, machine learning takes advantage of the available data to glean the insights and make the predictions hence the possibility of new insights hitherto unforeseen [19]. The target variable had a label which was binary, hence the supervised learning problem was classification in nature. Therefore, this paper attempts to address a classification problem that compared machine learning models that classify an insurance potential client as either taking up a policy or otherwise based on features of the potential customer, mostly social-demographic in nature. Whereas machine learning may not have been applied in insurance uptake prediction, it has had applications in other areas of insurance ranging from application to modelling of life insurance companies Krah et al. [11], as well as in export credit finance claims prediction [3].

2. Description of the Selected Machine Learning Models

For a sequence set of label $\mathcal{Y}$ and sequence domain set $\mathcal{X}$, we sought an optimal classifier h, that predicts new customer as taking up insurance or not such that the loss $Ls(h)$, in the test set is minimized. $h : \mathcal{X} \to \mathcal{Y}$. Where $\mathcal{H}$ is the hypothesis class that h is expected to fall [18].

$$h_S \in \underset{h \in H}{\operatorname{argmin}} \; L_S(h) \tag{1}$$

2.1. Logistic Regression Classifier

Logistic regression classifier is derived from linear regression but modified by a logistic function. For the outcome Y, which is an uptake of insurance or non-uptake for the current study, the outcome may be 1 or 0 otherwise.

$$Y = \begin{cases} 1, & = uptake \\ 0, & = \text{non uptake} \end{cases}$$

For $\mathbf{X} = x_1 + ... + x_n$ where $x_1 + ... + x_n$ are features under consideration. The probability that a customer or potential customer will take up insurance will be given by:

$$\pi(x) = E(Y|x_1....x_n) = \frac{exp\{\beta_0 + \beta_1 x_1 + ... + \beta_n\}}{1 + exp\{\beta_0 + \beta_1 x_1 + ... + \beta_n\}} \tag{2}$$

.

The cost function will then be given by:

$$cost(f(x), uptake) = \begin{cases} -logf(x), & \mathrm{i}fuptake = 1 \\ -log(1 - f(x), & \text{otherwise} \end{cases} \tag{3}$$

[18]

2.2. Support Vector machines (SVM)

SVM is a frontier hyperplane that optimally segregates two classes by seeking the largest margin between the nearest points of the training set of any class (referred to as support vectors). With the kernel trick, finite dimensional space features could be mapped into higher dimensional space hence making it possible to linearly separate despite the dimensional space [18]. SVM has been applied in diverse areas and has been found to have high accuracy in many instances including cancer diagnosis [10]. In this paper, the support vectors gave the largest margin between the insurance clients (and potential ones) who would uptake cover and those who would not uptake, based on the features.

2.3. Gaussian Naive Bayes (GNB)

GNB determines the posterior probability of taking up insurance of either the uptake or the non-uptake given the features. Given the value of the label, y, the algorithm takes Bayes's theorem but with the assumption of conditional independence between every pair of covariates.

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots x_n \mid y)}{P(x_1, \ldots, x_n)} \tag{4}$$

The y is the class label, $x_1, \ldots, x_n$ are the features. The equation 4 gives how the posterior distribution which is the hypothesis class label given the data is arrived at. Moreover, the equation 5 below gives the assumption of the Gaussian distribution of the covariates.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right) \tag{5}$$

The class label in this research was the uptake while the covariates were the features. The classifier therefore works by comparing the posterior distribution of each of the two classes (which may either be uptake or non-uptake) and takes the higher.

2.4. K Nearest Neighbour (KNN)

The classifier in KNN is such that for a training set S and for every point $\mathbf{x}\, in\, \mathcal{X}$, it outputs the highest of the label among $(y_{\pi i}(\mathbf{x}) : 1 \leq k)$, where $\pi_1(\mathbf{x}.....\pi_m(\mathbf{x})$ is the reordered set S, ordered as per their distance to $\mathbf{x}$. The distance in this case is Euclidean distance. The classifier finds the Euclidean distance between new data points for each training example. It then selects the K entries closest to the new data point. The label with the highest frequency in K entries will be the class label of the new data point. So if the most common is non-uptake, the new data point will be classified as non-uptake and vice versa [18].

2.5. Decision Trees (DT)

These are classifiers, $h\colon X \to Y$, that predict the label associated with an instance of, say variables, by moving from root node to a leaf. DT are built as branch-like fragments. The decision tree consists of root nodes, leaf nodes signifying the class labels, whereas the intermediate nodes denote non-leaf nodes. The data attribute with the highest priority in decision-making is selected as the root node. The splitting process of a decision tree is decided upon the data values of the respective nodes. The decision tree learns during the training phase and its effectiveness is evaluated during the testing phase. The depth and the distribution of training and test information of DT dynamically impact the performance and efficiency of the classifier [13]. In this research, DT were applied to predict insurance uptake; leaf nodes represented classification labels which may be either uptake or non-uptake.

2.6. Random Forest (RF)

RF is a tree based algorithm whose trees are assembled by bagging. The trees are independently trained. RF classifier use an ensemble of decision trees to predict insurance uptake based on features. The prediction is the outcome of sequential, binary decisions that are orthogonal splitting in the multivariate space of variables. In this case therefore, the RF classifier is taken as the classifying algorithm, that is a meta learner of several trees built independent of each other. RF has compared favourably with other ensemble decision tree-based models in previous studies. In some cases, it performs better than other learners [4] but in others like the case of [8] boosting performed better.

2.7. Gradient Boosting Machine and Extreme Gradient Boosting

Gradient boosting machines (GBM) and extreme gradient boosting (XGB) are tree based supervised learning models. Their ensemble method of learning is by boosting. The classification trees are sequentially trained to improve the performance of the next tree from the previous one. As a result, every new tree attempts to correct errors of the preceding tree. GBM and XGB only differ from the point that XGB use a more formalized regularization than GBM. These makes controlling for over-fitting better and gives better model performance. Gradient boosting methods have been found to perform well, yielding state of the earth results in most classification and regression tasks compared to other models [8].

3. Methodology

3.1. Data

The study used 2016 Kenya FinAccess household survey data. Among the main objectives of the survey is to measure access to and demand for financial services among adults. The sample of the survey was representative of the whole country and based on KNBS NASSEP V national household sampling frame. The survey undertook 10,008 interviews from 834 clusters across the country, with 14 households being targeted in each cluster. A KISH grid was then used to select respondents at the household level. The sample drawn was representative downwards to 13 sub-regions in the country which were clusters. These were: North Rift region, Central Rift region, South Rift region, Nyanza region, Western region, Eastern region, Coastal region, North Eastern region, Nairobi and Mombasa. The survey was intended to measure access to and demand for financial services among Kenyans aged 16 years and above. A nationally representative cross-sectional survey used a multi-stage stratified cluster sampling design. About 834 clusters were initially selected as primary sampling units (PSU), using probability proportional to size (PPS), from a national sampling frame. The Fifth National Sample Survey and Evaluation Program (NASSEPV) was designed by the Kenya National Bureau of Statistics, according to Kenya's previous population census (2009 population census). Furthermore, there was stratification according to urban and rural areas together with the country's 47 counties hence resulted in 92 strata. The second stage involved selecting 14 households in each cluster. In the final stage, one individual aged 16 years old and above, was randomly selected per household using the KISH grid. 8,665 interviews in 820 clusters were conducted. One person was interviewed per household. Data was collected on socio demographic characteristics, access, and use of financial services including mobile money, and social health insurance enrolment [6].

3.2. Features Selection

The aim of the supervised learning adopted in this paper was to find a learner that classifies an insurance customer, or potential one in an optimal and robust manner. The data set contained over 100 variables but only 30 social demographic features were selected. Those that were highly co-varied and those with the least correlation were picked. This reduced the dimensionality hence reduced likely-hood of overfitting. From the 30 features, 18 features were selected based on univariate selection where features with the strongest relationship with the label were taken. The univariate compares each feature with the label to check if there is a significant relationship. This feature selection method mechanism is such that it selects the best features based on univariate statistical tests. It compared each feature to the label variable, which was uptake in the case of this research, to check whether there was any statistically significant relationship between them. When comparing the relationship between one feature and the target variable, we ignored the other features. Each feature therefore had a test score. Lastly, the test scores from each feature were compared, and the features with top scores were selected. This method was chosen for its flexibility for the given kind of data.

The label set was uptake, while the instances in the domain set considered included vector features: gender, marital status, age group, education, numeracy, pace of residence, internet access, own phone, electricity as the main light source, smartphone, youth, wealth quintile, having a bank product, top trusted provider, second top trusted provider, residence, household size, and having some fund set aside for emergency and the sub-region that one lives within the country that the respondent resided. One hot encoding was performed on the categorical features to enable better prediction as factors. The uptake in this study implies the uptake of insurance regardless of the line or class of insurance. Those included in this label are those who had any kind of insurance cover; whether life, non-life or medical.

3.3. Handling Class Imbalance

Class imbalance problem arises when the data has a proportion of one class (majority class) being significantly higher in ratio than another (minority class). This could be alleviated by various techniques including oversampling of the minority class or undersampling of the majority class to balance both classes [2]. The data set used in the study, as is commonly the case in real-world settings, was imbalanced to an extent proportionally. Figure 1 shows the data balance proportion for the data used. The proportion of those who did not have insurance to those who had was 6807:1858 which is 3.66: 1 hence the data was unbalanced with the minority class being 21% of the data while the majority class is 79%. The 1 demonstrates class to some level imbalance between uptake (the minority class), and non-uptake (the majority class). When the event of interest is underrepresented uptake (the minority class), the majority class (non-uptake), there is a tendency to hinder the classification accuracy. Data imbalanced was handled by both up-sampling and down-sampling.

3.4. Model Performance Measures

Precision scores refer to the number of true positives divided by all positive predictions. Precision is also called Positive Predictive. Value is a measure of a classifier's exactness. Low precision indicates a high number of false positives. Recall score refers to the number of true positives divided by the number of positive values in the test data. Low recall indicates a high number of false negatives. Sensitivity or the True Positive Rate is a measure of a classifier's completeness. F1-score is the weighted average of precision and recall.

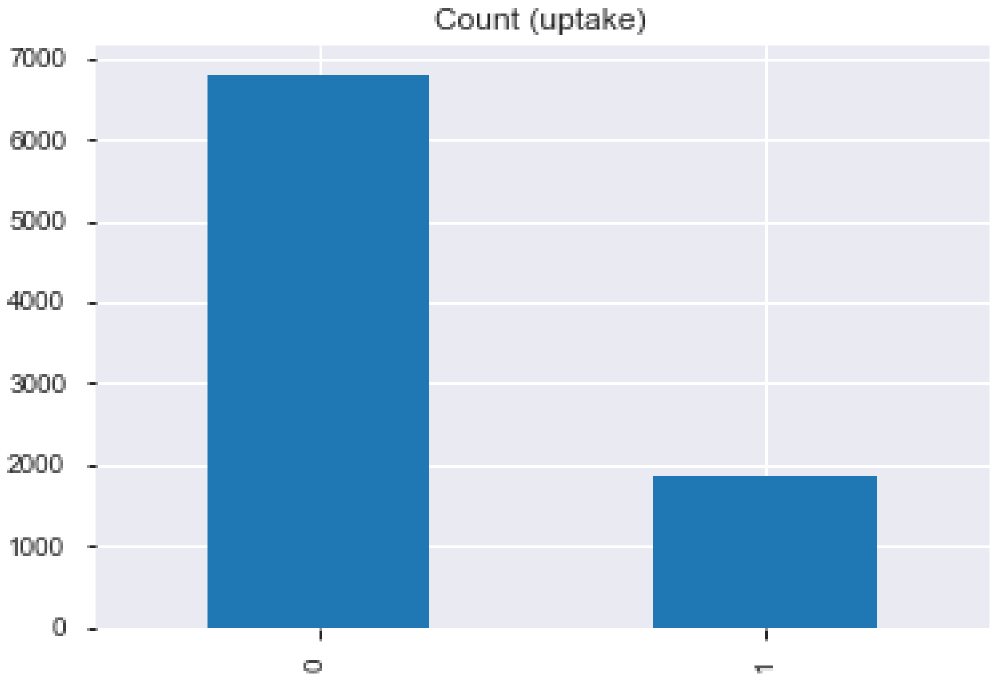

Figure 1. Data Imbalance

Confusion Matrix is a table showing correct predictions and types of incorrect predictions. True positives (TP) refers to the number of cases correctly identified as having taken insurance False positives (FP) are the number of cases incorrectly identified as having taken insurance True negatives (TN) are the number of cases correctly identified as not having insurance False negatives (FN) are the number of cases incorrectly identified as not having insurance.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{6}$$

$$Sensitivity = \frac{TP}{(TP + FN)} \tag{7}$$

$$Specificity = \frac{TN}{(TN + FP)} \tag{8}$$

3.5. Hyper-parameter optimization

Grid searches were conducted to find hyper-parameters that would yield optimal model performance. A 5-fold cross-validation technique was used based on accuracy as an evaluation metric. Table 1 shows the hyper-parameters tuned in RF. Hyper-parameters for the RF that were tuned were: n_estimators, max_features, min_samples _split and bootstrap. The n_estimators which represent the number of trees in the forest were optimized between 80 and 150 with a range of 10.

Table 1. Hyper-parameter Optimization for RF

Parameter	Range	Optimal Value
n_estimators	[80 to 150, interval of 10]	110
max_features	[auto, sqrt, log2]	auto
min_samples_split	[2,4,6,8]	2
Bootstrap	[True, False]	True

Usually, the higher the n_estimators, the better the accuracy, but on optimizing an optimum of 110 was found. The max_features is the maximum number of features that the model considers when splitting a node. The search was among auto, sqrt and log2. The grid search for max_features was optimized at auto. The min_samples _split represents the minimum number of data points, from the total, that should be placed in a node before the node is split. The search was between 2,4,6 and 8 and it was optimized at 2. Finally, on bootstrap which represents the method used by the model to sample the data points; whether with replacement or without. In this case it is optimized at true implying with replacement.

Table 2. Hyper-parameter Optimization for GBM

Parameter	Range	Optimal Value
n_estimators	[100,200,300]	100
min_samples learning_rate	[0.01, 0.02, 0.05, 0.1]	0.05

Table 2 shows the hyper-parameters tuned in GBM. For the GBM, n_estimators were tuned among 100,200 and 300 and it optimized at 200. The parameter learning_rate was to control the weighting of new trees added to the model. The parameter was tuned between 0.01, 0.02, 0.05 and 0.1. The optimal learning_rate was found to be 0.05.

Table 3. Hyper-parameter Optimization for XGB.

Parameter	Range	Optimal Value
n_estimators	[500 to 1500]	1000
max_depth	[auto, sqrt, log2]	auto
max_features	[0.2 to 1]	0.9
gamma	[0.1 to 1]	0.1

Table 3 shows the hyper-parameter tuning on XGB. The parameters that were optimized were: n_estimators, max_depth, max_features and gamma. On n_estimators, they were tuned between 500 and 1500 and the optimal was found at 1000. For max_depth, it was tuned among auto, sqrt and log2 and the optimal arrived at was auto. On the other hand, max_features were tuned between 0.2 and 1 and optimal arrived at was 0.9. Finally, gamma tuned in at 0.1.

4. Results and Discussion

Python version 3.6.1 was used as a tool for analysis. Python was used because it is malleable. Moreover, the various libraries that python has made the analysis easier. Jupyter notebook was the environment of choice for its simple interface. The results shown are for test sample in each case. After data cleaning and selection of relevant features data was split into three sets: training, validation and test sets in the ratio of 0.75:0.15:0.10. For the training set 75% of the data were used to train the algorithms. For validation set 15% of the data were held back from training of the model and were used to give an unbiased sagacity of model efficiency. The validation set was used to evaluate performance on data which were unseen when test data was locked away. For the test set 10% of the data were held back from training of the model and were used to give an unbiased sagacity of a final model efficiency. The test set was locked away till fine-tuning of the model was complete thereafter an unbiased evaluation of the final models were obtained. Pawluszek-Filipiak and Borkowski [15] observed that performance metrics of the models, F1 score and Overall Accuracy, decreased as the train-test ratio decreased. This implies that as the training sample decreases the performance metrics tend to decrease. In line with models' need substantial data to train upon, Poria et al. [17] used 80% of the data as set training and the remaining 20% was partitioned equally between validation and test. As a result there was the need to choose a train-validation-test split ratio that not only optimized the accuracy but also adequately measured the extent to which the model would perform in "unseen" data. These considerations informed the choice of the train-validation-test split ratio that was used.

4.1. Comparison on Unbalanced data

Table 4. Machine learning model Performance for the unbalanced data.

Number	Model	Precision	Recall	F1_scores	Accuracy
0	Logistic	0.6855	0.5430	0.6060	0.8485
1	GNB	0.4576	0.7258	0.5613	0.7565
2	RF	0.6301	0.3907	0.4823	0.8200
3	DT	0.4339	0.4821	0.4567	0.7538
4	SVM	0.7265	0.4857	0.5822	0.8504
5	KNN	0.5781	0.4910	0.5310	0.8138
6	GBM	0.7054	0.5108	0.5925	0.8492
7	XGB	0.7204	0.5126	0.5990	0.8527

Table 4 shows machine learning models and their respective precision scores, recall scores, F1_scores and accuracy on the real unbalanced data. The XGB, SVM, GBM and Logistic Regression classifiers have the highest accuracy (0.85) followed by RF (0.82), KNN (0.81), GNB (0.76) and the lowest DT (0.75). For F1_scores, logistic regression shows highest (0.61), then XGB (0.60), GBM (0.59), SVM (0.58), GNB (0.56), KNN (0.53), RF (0.48), and DT has the lowest (0.46). Despite all the accuracy scores of each of these being at least 0.75, their respective precision, recall and F1_scores are all below 0.75. This may be an indicator of some skew in the data hence the need for data balancing before training.

The machine learning models improved their skill on unseen data upon the cross validation. All the models improved their respective precision scores. For logistic regression from 0.69 to 0.77, GNB from 0.46 to 0.69, RF from 0.63 to 0.76, DT from 0.43 to 0.65, SVM from 0.73 to 0.77, KNN from 0.58 to 0.75, GBM from 0.71 to 0.7720, and XGB from 0.72 to 0.76. Similarly, there was improvement in all recall scores. For logistic from 0.54 to 0.72, GNB from 0.73 to 0.75, RF from 0.39 to 0.6748, DT from 0.4821 to 0.6493, SVM from 0.49 to 0.68, KNN from 0.49 to 0.72, GBM from 0.51 to 0.71 and XGB from 0.51 to 0.69. The f1-score improvements were: logistic regression from 0.61 to 0.74, GNB from 0.56 to 0.71, RF from 0.48 to 0.70, DT from 0.46 to 0.65, SVM from 0.58 to 0.71, KNN from 0.53 to 0.73, GBM from 0.59 to 0.73 and XGB from 0.60 to 0.72. There was a slight improvement in accuracy for most of the machine learning models except for logistic regression and SVM. The respective changes in accuracy after the cross validation were: logistic regression from 0.85 to 0.84 GNB from 0.76 to 0.77, RF from 0.82 to 0.83, DT from 0.75 to 0.77, SVM from 0.85 to 0.84, KNN from 0.81 to 0.83, GBM from 0.85 to 0.84, and XGB from 0.85 to 0.84. There were not many improvements in the respective accuracy because the data was still unbalanced. This implies with cross validation the accuracy generally remained the same but the f1-scores generally rose with the k-fold cross-validation.

4.2. Comparison on Balanced data

Table 5. Machine learning models Performance unbalanced data with cross validation.

Number	Model	Precision	Recall	F1_scores	Accuracy
0	Logistic	0.7726	0.7228	0.7423	0.8442
1	GNB	0.6912	0.7507	0.7059	0.7712
2	RF	0.7456	0.6748	0.6977	0.8269
3	DT	0.6474	0.6493	0.6483	0.7654
4	SVM	0.7704	0.6809	0.7080	0.8365
5	KNN	0.7495	0.7155	0.7297	0.8327
6	GBM	0.7720	0.7115	0.7338	0.8423
7	XGB	0.7630	0.6944	0.7182	0.8366

Table 5 shows machine learning models and their respective precision scores, recall scores, F1_scores and accuracy on the real unbalanced data but with cross validation test_size=0.1, validation_score = 0.15. Stratification was with respect to y which is uptake in the case, with a set the random_for reproducibility. This stratifies the parameter making a split so that the proportion of values in the sample produced will be the same as the proportion of values provided to the parameter stratify; it ensures that the in cross validation, the skews within the folds are similar. High accuracy is observed among logistic, GBM, XGB and SVM (0.84) then KNN and RF (0.83) and finally GNB and DT (0.77). On the other hand for F1_scores logistic had highest (0.74), GBM and KNN (0.73), XGBM (0.72), SVM and GNB (0.71), RF (0.70) and finally DT (0.65).

Table 6. Machine learning model Performance for the upsampled data.

Number	Model	Precision	Recall	F1_scores	Accuracy
0	Logistic	0.7775	0.7776	0.7775	0.7775
1	GNB	0.7440	0.7436	0.7432	0.7433
2	RF	0.9493	0.9467	0.9462	0.9462
3	DT	0.9311	0.9250	0.9240	0.9242
4	SVM	0.8193	0.8192	0.8191	0.8191
5	KNN	0.8328	0.8250	0.8231	0.8240
6	GBM	0.7921	0.7922	0.7922	0.7922
7	XGB	0.7874	0.7874	0.7873	0.7873

Table 6 shows machine learning models and their respective precision scores, recall scores, F1_scores and accuracy on the upsampled data. RF leads with the highest accuracy of 0.95 followed by DT(0.92), KNN (0.82), SVM (0.82), GBM, XGB (0.79), logistic regression (0.78) and lastly GNB (0.74). However, upon hyper-parameter tuning, GBM and XGB raise their respective accuracy. Nevertheless, RF showed the highest precision score, recall score, F1_score and accuracy hence can be taken as the optimal model in this instance. Here, RF are more robust than other classifiers. This could be explained by its being an ensemble algorithm. The findings are in corroboration with Han et al. [9] who assert that ensemble algorithms tend to perform better than stand-alone algorithms. However, GBM and XGB gave lower accuracy than the DT classifier unlike expectation. Hence, we could conclude that for this kind of upsampled data, ensemble trees by bagging tends to perform better than by boosting for the kind of data under this study.

Table 7 shows machine learning models and their respective precision scores, recall scores, F1_scores and accuracy on the real down sampled data. The highest accuracy score was observed in XGB (0.87), then GBM (0.86), logistic regression (0.83), KNN, GNB and RF (0.82) and Finally decision tree (0.72). For F1_scores, XGB has (0.87), then GBM (0.86), logistic (0.83), KNN, GNB and RF (0.82) and Finally decision tree (0.72). In the case of downsampled data XGB and GBM showed higher accuracy than other models (0.87 and 0.86 respectively). Both of them are tree based ensemble learners which are assembled by boosting. This allows us to construe that for the kind of data used, tree-based learners ensemble by boosting are more robust than others. This corroborates with Golden et al. [8] who have found GBM to perform better than other algorithms.

Despite being the same data source, the learners had different metrics for upsampled and downsampled data. The accuracy for the learners when the data was upsampled, vis a vi when

Table 7. Machine learning model Performance for the down sampled data.

Number	Model	Precision	Recall	F1_scores	Accuracy
0	Logistic	0.8292	0.8314	0.8297	0.8303
1	GNB	0.8200	0.8216	0.8205	0.8214
2	RF	0.8207	0.8232	0.8219	0.8214
3	DT	0.7210	0.7202	0.7205	0.7232
4	SVM	0.8125	0.8150	0.8121	0.8125
5	KNN	0.8207	0.8232	0.8209	0.8214
6	GBM	0.8558	0.8576	0.8564	0.8571
7	XGB	0.8649	0.8674	0.8655	0.8661

downsampled were: logistic (0.78, 0.83), GNB (0.74, 0.82), RF (0.95, 0.82), DT (0.92, 0,72), SVM (0.82, 0.81), KNN (0.82, 0.82), GBM (0.79, 0.86), XGB (0.78, 0.87). This could imply that when the data is downsampled learners presume different distributions from when upsampled, despite being the same data. However, SVM and KNN do not seem to show remarkable differences in accuracy when the data is either downsampled or upsampled.

4.3. Area under Receiver Operating Characteristic Curves (AUCs) and Confusion Matrices

Table 8. Confusion Matrices

Number	Model	TP	TN	FP	FN
0	Logistic	164	154	40	51
1	GNB	159	191	45	14
2	RF	190	201	14	4
3	DT	170	204	34	1
4	SVM	158	191	46	14
5	KNN	159	191	45	14
6	GBM	169	166	35	37
7	XGB	179	190	25	15

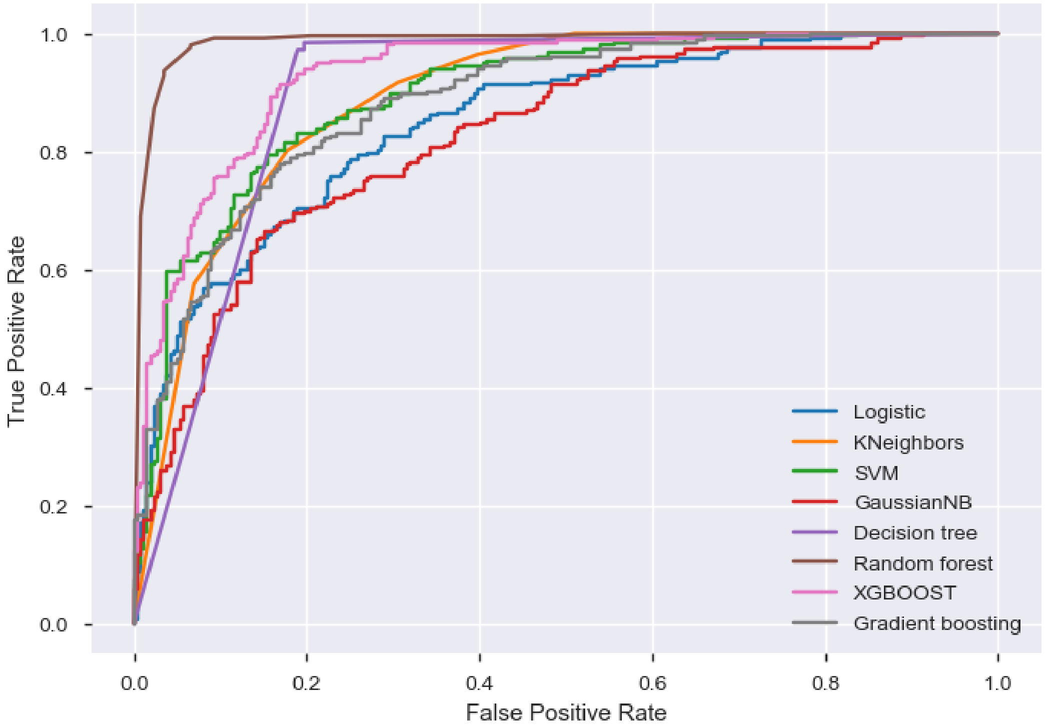

Figure 2. Areas under receiver operating characteristics curves (AUCs)

Upon imputing the optimized hyper-parameters, the models were retrained and AUCs and confusion matrices for the various models were drawn. Table 8 shows values of true positives TP, TN, FP and FN that were extracted from the confusion matrices for various models. In giving TP, RF leads (190), then XGB (179), DT (170), GBM (169), Logistic regression (164), GNB (159) and KNN gives the least (150). This means that it is more likely for a RF model to predict that one would take up insurance cover and that person does take up cover compared to other models. In TN, DT has the highest (204), then RF (201), GNB, KNN (191), XGB (190), GBM (168) and lastly (154). In giving FN, DT has the lowest (1), then RF (4), KNN (14), GNB (14), XGB (15), GBM (37) and logistic regression (51). On FP, RF had the lowest (14), followed by XGB (25), decision tree (34), GBM (35), logistic regression (40) lastly KNN and GNB (both 45). This implies that RF are least likely to make type II errors in predicting uptake compared with other classifiers. RF seems to be the most robust since it has the highest true positives and the least false positives. Nevertheless, other tree based classifiers seem to do well in the data.

Moreover, Figure 2 shows areas under receiver operating characteristics curves (AUCs) for the various models. The AUCs under the various models were: 0.8481 for logistic regression, 0.8914 for K Nearest Neighbors classifier, 0.8220 for GNB, 0.8940 for SVM, 0.8962 for decision tree,0.9866 for RF, 0.9300 for XGB (referred to as XGBOOST on the figure), and 0.8823 for GBM. Based on AUCs, RF performed best compared to all other models since it gave the highest area under receiver operating characteristics curve, followed by XGB classifier. This corroborates Blanco et al. [4] who found RF to be a stronger model in prediction of the efficiency of the fullerene derivatives-based kind of ternary organic solar cells. This also implies that tree-based models tend to perform better than others for this kind of data, since both RF and XGB are tree-based models.

4.4. Feature importance

Feature importance in this study has been employed to get an understanding into how the features contribute to model predictions. As previously proposed by Casalicchio et al. [5], the effect of features, their respective contributions as well as their respective attributions; describe how and, to some level, the extent to which each feature contributes to the prediction of the model. Further, as Pesantez-Narvaez et al. [16] adds, the contribution of each feature to the final outcome as given by the feature importance is based on Gini impurity. Feature importance was taken in order to identify important features that contribute to the greatest extent to the prediction of uptake. This enabled analysis and comparison of the feature importance across various observations in the sampled dataset. The feature importance was obtained from the RF model and since RF is a tree based model, it gives the extent to which each feature contributes to reducing the weighted Gini impurity.

Table 9. Feature importance.

Rank	Feature	Importance
1	Having a bank product	0.191
2	Wealth Quintile	0.111
3	Sub Region	0.109
4	Education	0.088
5	Age Group	0.068
6	Most Trusted provider	0.051
7	Nature of residence	0.050
8	Numeracy	0.048
9	Household size	0.047
10	Marital status	0.041
11	2nd Most Trusted provider	0.039
12	Ownership of a phone	0.038
13	Having a set Emergency fund	0.033
14	Having electricity as a light source	0.031
15	Gender	0.030
16	Urban vs Rural	0.026
17	Being a youth	0.026
18	Having a smartphone	0.023

Based on the AUCs and Confusion Matrices, tree-based models seemed to be more robust in uptake prediction. Moreover, RF showed the highest AUC and hence this model was used to extract the importance of each feature in the uptake prediction. Table 9 shows the feature importance from the RF model in predicting the insurance uptake. All the features show non-zero importance but their rank from the most important to the least important is having a bank product, wealth quintile, subregion, level of education, age, group, most trusted provider, nature of residence, numeracy, household size, marital status, second most trusted provider, ownership of a phone, having a set emergency fund, having electricity as a light source, gender, nature of residential area whether it is urban or rural, being a youth, and having a smartphone.

The result suggests that the most important factor is whether one has a bank product or not. This implies that individuals who have a bank product tend to have higher uptake of insurance compared to those who do not. This could also imply that many individuals who had a bank product also had an insurance product. The second most important feature is wealth quintile. This implies that the material wealth of an individual plays critical role since the wealthier and individual the higher would be the ability to pay for the insurance premiums. The ability to pay is a great factor in determining uptake. The subregion being the third factor could be construed to imply that the insurance products are not evenly distributed nationally. Interestingly, being a youth and having a smartphone did not show much importance in determining uptake. This is despite much of the population being young. This could imply that the insurance products in the market are not appealing to the youths, or they could be too expensive for them. More could be done on product engineering to make insurance products more affordable and more appealing so that more of the young populace could benefit from insurance.

5. Conclusions and Recommendations

It could be construed that with the unbalanced data, performance was lower compared to performance in the balanced data. This could mean that the data imbalance problem is a significant contributor to poor model performance in insurance uptake prediction. Moreover, the learning metrics improved when the data was balanced by either upsampling the minority class (uptake of insurance for the case of data used) or downsampling majority class (non-uptake of insurance for the case of data used). Therefore, alleviating data imbalance results in a more robust model in insurance uptake prediction.

In the study, ensemble learners mostly tended to perform better than stand-alone algorithms. For upsampled data, RF, which is assembled by bagging, performed better than other machine learning classifiers that were considered. RF had the highest accuracy of 0.95. But for downsampled data GBM and XGB, which are assembled by boosting, did better than the others. GBM and XGB had accuracy of 0.87 and 0.86 respectively. Therefore, ensemble learners could be said to be the most robust for this kind of data. Moreover, RF, GBM and XGB are all tree based models, therefore tree based ensemble machine learning models could be said to be robust for insurance uptake prediction.

Despite being the same data source, the learners had different metrics for upsampled and downsampled data. It could therefore be concluded that when the data was downsampled learners presume that they were drawn to different distributions from when they were upsampled, despite being the same data. However, SVM and KNN did not seem to show remarkable differences in accuracy when the data is either downsampled or upsampled. Further study could be done to find out if this lack of remarkable difference came by chance or if it stems from the nature of the classifiers. The most important feature in predicting uptake is having a bank product. This could imply that bancassurance is a viable channel or distribution of insurance products since the banked population are more likely to take up insurance. The wealth quintile was the second most important factor in the rank of feature importance. This therefore calls for insurance providers to come up with innovative products that would be affordable to the majority of the population. Spatial characteristics were the third most important factor. This could imply the distribution of insurance is not even in the nation. A further look at this could be done with multilevel modelling to establish the extent of the different levels of variation in the data.

In recommending improvement from here, we suggest studies be done on specific lines of insurance with machine learning models that we have herein found to be most robust, particularly Random Forests.

6. Acknowledgment

We thank the African Center of Excellence in Data Science and the University of Rwanda for partly funding this study. We also appreciate FinAcess, FSD and KNBS for publicly providing the data used in this study.

References

1. AKI. 2015. Insurance Industry Annual Report 2015. Technical report, Association of Kenya Insurers.
2. Amin, Adnan, Sajid Anwar, Awais Adnan, Muhammad Nawaz, Newton Howard, Junaid Qadir, Ahmad Hawalah, and Amir Hussain. 2016. Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. *IEEE Access 4*, 7940–7957.
3. Bärtl, Mathias and Simone Krummaker. 2020. Prediction of claims in export credit finance: A comparison of four machine learning techniques. *Risks 8*(1), 22.
4. Blanco, Carlos M Guio, Victor M Brito Gomez, Patricio Crespo, and Mareike Ließ. 2018. Spatial prediction of soil water retention in a páramo landscape: Methodological insight into machine learning using random forest. *Geoderma 316*, 100–114.
5. Casalicchio, Giuseppe, Christoph Molnar, and Bernd Bischl. 2018. Visualizing the feature importance for black box models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 655–670. Springer.
6. Central Bank of Kenya and FSD Kenya and Kenya National Bureau of Statistics. 2016. FinAccess Household Survey 2015. doi:10.7910/DVN/QUTLO2.
7. Commission, Africa Union et al.. 2017. Agenda2063-the africa we want.
8. Golden, Chase E, Michael J Rothrock Jr, and Abhinav Mishra. 2019. Comparison between random forest and gradient boosting machine methods for predicting listeria spp. prevalence in the environment of pastured poultry farms. *Food Research International 122*, 47–55.
9. Han, Taihao, Ashfia Siddique, Kamal Khayat, Jie Huang, and Aditya Kumar. 2020. An ensemble machine learning approach for prediction and optimization of modulus of elasticity of recycled aggregate concrete. *Construction and Building Materials 244*, 118271.
10. Huang, Shujun, Nianguang Cai, Pedro Penzuti Pacheco, Shavira Narrandes, Yang Wang, and Wayne Xu. 2018. Applications of support vector machine (svm) learning in cancer genomics. *Cancer Genomics-Proteomics 15*(1), 41–51.
11. Krah, Anne-Sophie, Zoran Nikolić, and Ralf Korn. 2020. Machine learning in least-squares monte carlo proxy modeling of life insurance companies. *Risks 8*(1), 21.
12. Lambregts, Timo R and Frederik T Schut. 2019. A systematic review of the reasons for low uptake of long-term care insurance and life annuities: Could integrated products counter them?
13. Naganandhini, S and P Shanmugavadivu. 2019. Effective diagnosis of alzheimer's disease using modified decision tree classifier. *Procedia Computer Science 165*, 548–555.
14. Olayungbo, DO and AE Akinlo. 2016. Insurance penetration and economic growth in africa: Dynamic effects analysis using bayesian tvp-var approach. *Cogent Economics & Finance 4*(1), 1150390.
15. Pawluszek-Filipiak, Kamila and Andrzej Borkowski. 2020. On the importance of train–test split ratio of datasets in automatic landslide detection by supervised classification. *Remote Sensing 12*(18), 3054.
16. Pesantez-Narvaez, Jessica, Montserrat Guillen, and Manuela Alcañiz. 2019. Predicting motor insurance claims using telematics data—xgboost versus logistic regression. *Risks 7*(2), 70.
17. Poria, Soujanya, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 873–883.
18. Shalev-Shwartz, Shai and Shai Ben-David. 2014. *Understanding machine learning: From theory to algorithms*. Cambridge university press.
19. Venderley, Jordan, Vedika Khemani, and Eun-Ah Kim. 2018. Machine learning out-of-equilibrium phases of matter. *Physical review letters 120*(25), 257204.
20. Wrede, Xavier Giné Bernardo Ribeiro Peter. 2019. *Beyond the S-curve: Insurance Penetration, Institutional Quality and Financial Market Development*. The World Bank. doi:10.1596/1813-9450-8925.