International Journal of Computer (IJC)
ISSN 2307-4523 (Print & Online)
http://ijcjournal.org/

Predictive Model for Computing Health Insurance Premium Rates Using Machine Learning Algorithms

Angela D. Kafuria*

African Centre of Excellence for Data Science, University of Rwanda, KK 737 St, Kigali, Rwanda

Abstract

The health care systems depend heavily on out-of-pocket payments, the mechanism that is a barrier to universal health coverage, as it contributes to inefficiency, inequity and cost. To solve this challenge, people are encouraged to enrol on health insurance schemes to reduce the burden of out-of-pocket payments. There is a strong need for insurance companies to develop models that accurately predict medical expenses for the insured population. The variables; Age, sex, body mass index, number of children and region attributes were used to formulate a predictive model to determine health insurance charges using different Machine learning algorithms techniques. The findings showed that the following variables were significant; age (p = 0.000), BMI (p = 0.001), smoking (p = 0.000) and region (0.046). Therefore, these attributes can be said to be the determinants of health insurance charges. Five (5) models that were used in predictive analysis were evaluated. These models were Multiple Linear Regression (MLR), K-nearest Neighbors (KNN), Least Absolute Shrinkage and Selection Operator (LASSO), Extreme Gradient Boosting (Xgboosting) and Random Forest Regression (RFR) The models' performance evaluation findings indicated Gradient Boosting and RFR were the best models in prediction with the following values R2 = 85.5%, MAE = 2688.2, RMSE = 4748.7 and R2 = 85.3%, MAE = 2726.4, RMSE = 4783.8 respectively. The insurance companies that seek to develop a model for prediction premiums are recommended to use Extreme Gradient Boosting and RFR for a more accurate model.

Keywords: Health Insurance; Insurance premiums; Machine learning; Predictive models.

1. Introduction

Health care systems in developing countries depend heavily on out-of-pocket payments, the mechanism that is a barrier to universal health coverage, as it contributes to inefficiency, inequity and cost [1]. Health insurance, which is coverage against the risk of incurring medical and related financial costs, is one of the ways that people in various countries pay for their medical needs [2].

It is a mechanism that can be used to ensure people have good health and well-being. This is a priority to every human being, hence access to health care coverage is worldwide priority.

---

* Corresponding author.

But due to the high rates that are charged by insurance companies, many people are without health insurance and so fail to access timely health services which results in high death rates. The author [3] suggested that one way to encourage healthcare insurance enrolment is to charge rates that are affordable for many people and that give quality service to the clients. However, health insurance rates calculations are often complex as they need to determine the rates that are acceptable by both insurance companies and beneficiaries; Insurance companies need to make money by collecting more money than they spend on the medical expenses of their beneficiaries, hence make a profit and continue to stay in businesses. These companies price the premiums based on the probability of certain events occurring among a pool of people [4]. However, the medical expenses and other associated costs are difficult to estimate because the costliest conditions are rare and seemingly random [5]. Another complex part of estimating medical expenses is that the occurrence of certain diseases differs from one person to another and from one segment of the population to the other. Therefore, there is a need for a fair premium calculation model that suits the unique population factors. In this regard, this study used demographic and behavioural data from the patients to develop a predictive model. While previous studies used conventional statistical methods, this study used machine learning logarithms to develop a predictive model. It compares the performance of several models to find the most suitable

2. Literature Review

Health insurance combines two concepts, health concepts and insurance concepts. WHO defines health as a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity [6]. However, this definition of health has been challenged as being vague by the article by [7] that instead introduced a new concept of health as the ability to adapt and to self-manage, in the face of social, physical and emotional challenges. On the other hand, the term insurance refers to a method that households and firms use to prevent any single event from having a significant detrimental financial effect [4]. In a legal context, a contract of insurance is that whereby one party, the insurer, undertakes, for a premium or an assessment, to make a payment to another party, the policyholder or a third party, if an event that is the object of risk occurs [8].

In every country, some people are unable to pay directly or out of pocket for the healthcare services they need, or financially they may be seriously disadvantaged by doing so [2]. Thus, health insurance is very important as it ensures universal health care coverage. It is recommended as one of the best mechanisms to ensure better health and wellbeing of people. It is also a very useful practice in low-income communities as it protects insured persons from paying high treatment costs in the event of sickness by covering medical expenses that arise due to an illness [9]. These expenses could be related to the cost of medicines, doctor consultation fees or hospitalization costs. The purchase of health insurance reduces the risks and unpredictability inherent in a consumer's health care expenses. The consumer pays for a health insurance policy and then is subsequently (partly) reimbursed for his or her future expenditures on health care [10].

Previous studies evaluated predictive regression models on health insurance using demographic data as well as behavioural data. The authors [11] conducted a predicate analysis on the medical health insurance cost based on gender age, smoking habit, body mass index (BMI), number of children and region, using the medical information and costs. Using machine learning techniques, the study applied four regression models to the

dataset; Multiple Linear Regression, Support Vector Regression, Decision Tree Regression and Random Forest regression. The study results indicated that, among the four algorithms, Random Forest Regression gives better results. Also, age and BIM were found to have a strong influence on medical insurance charges.

A study by [12] forecasted health insurance premium charges based on age, sex, BIM, number of kids, smoking and region of the person living, medical cost personal data. The study used nine deep neutral networks and machine learning regression models; Multiple Linear Regression, Generalized Additive Model, Support Vector Machine, Random Forest Regressor, Classification and Regression Trees, XGBoost, k-Nearest Neighbors, Stochastic Gradient Boosting, and Deep Neural Network. The findings showed that Stochastic Gradient Boosting offered the best efficiency.

Author [13] predicted the insurance premium charge based on age, BMI, smoking, and the number of children. The study applied three machine learning techniques; multiple linear regression, random forest and Neural Network. The findings showed that the neural network did a better job of predicting insurance charges. Regarding the premium determinants, smoking was found to have highest influence on health insurance charges, followed by BMI and age.

A study by [14] used many attributes including sex, wealth quintile, region, education level, age group, household size, marital status, ownership of a phone, ownership of smartphone, most trusted providers, nature of residence, numeracy, having a set of an emergency fund, having electricity as a light source, having a bank product, urban versus rural and being a youth. The study compared the performance of seven (7) machine learning models; Logistic Regression Classifier Logistic, Support Vector machines (SVM), Gaussian Naive Bayes (GNB), K-Nearest Neighbor (KNN), Decision Trees (DT), Random Forest Regression (RFR), Gradient Boosting Machine and Extreme Gradient Boosting). Two models were found to show highest accuracy and precision in prediction. Regarding the attributes, 'having a bank product', wealth quintile, region a person is living and education level had significant influence on premium charges.

Another study that used predictive machine learning models to forecast expenditures, especially for the high-cost high-need patients was the study by [15] that used multiple features including demographic variables (age, sex, race/ethnicity and disabled status), diagnoses, medical procedures and medications. In this study four predictive models were applied; ordinary least squares linear regression, Least Absolute Shrinkage and Selection Operator (LASSO), gradient boosting machine (GBM), and recurrent neural networks (RNN). The LASSO and GBM were found to be more effective in generating interpretable contributions and finding important input variables than the RNN model.

In [16] used age, family size, region a person is living (county), maximum pocket and metal level features to evaluate .four regression models; Multiple Linear Regression, Decision tree Regression, AdaBoost Regression, and Gradient Boosting Decision Tree Regression. The study found that the Adaboost model which is built upon a decision tree is the best performing models. Regarding the features with significant influence, the findings indicated that family size and age had significant effects on premiums. The maximum out of pocket and deductible showed a negative sign, indicating that these variables negatively correlated with premiums.

Based on the studies above, several factors were identified as determinants of premium rates payment (Table 1). Occupation is added to the list because it was presented by [17] as an indirect factor.

Table 1: Factors affecting premium payment according to previous studies

Direct factors	Indirect factors
Age	Occupation
Sex	Economic status
BMI	Smoking
Past medical history	Types of plans chosen
Education level	Region a person residence
	Number of children/Family size

Among the variables presented above (Table 2), six (6) variables were selected to be used in this study; smoking, age, sex, BMI, number of children and region a person lives as the independent variable while health premiums paid by a person who is insured as the dependent variable.

3. Predictive Modelling

Multiple regression relies on different algorithms to solve various data problems. The multiple regression algorithms process data to learn the related patterns about individuals, business processes, transactions, events, and so on [18]. Machine learning algorithms build a model based on the "training data", to make predictions or decisions without being explicitly programmed to do so. The model is trained from historical data and the outcome is generated for the new test data. It involves two phases; Model training and Model testing. Training data contain input and target values then the algorithm picks up the pattern and maps the input values to the output and uses it to predict [18,19].

3.1 Multiple Linear Regression (MLR)

Multiple linear regression (MLR) is an extension of simple linear regression where there is one dependent variable (Y) and two or more independent variables (x_1, x_2, x_3, … x_n). The independent variable(s) can be continuous or discrete, while dependent variable is continuous. For this study, the value of x (independent variables) are as follows; x_1= Age, x_2 = Sex, x_3 = BMI, x_4 = children, x_5 = smoker, x_6 = region

$$\text{Insurance charges (y)} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Sex} + \beta_3 \text{BMI} + \beta_4 \text{Children} + \beta_5 \text{Smoker} + \beta_6 \text{Region} \quad (1)$$

In this dataset, the dependent variable is medical charges and independent variables are age, gender, smoker, BMI, children, and region. MLR uses the ordinary least-squares (OLS) method to find a best-fitting line that involves multiple independent variables [11]. Most researchers have been using a generalized linear model (GLM) for the health premium prediction because of its simple interpretability of the fitted parameters. This study used a supervised learning technique under machine learning. This is because with supervised ML there is a more accurate model compared to GLM.

3.2 K-nearest Neighbours (KNN)

KNN is non-generalizing learning. Instead of constructing a general internal model, it stores all instances corresponding to training data in n-dimensional space. Here, the two parameters considered are the value of K which is a parameter that refer to the number of nearest neighbours (in our case we used 10 neighbours) and the distance function whereby the distance between the new point and each training point is calculated, and then the closest points are picked. There are various methods of calculating distance, the common ones are three; Euclidean distance, Manhattan distance and Hamming distance. Euclidean distance (D_H) is the squire root of the sum of the squired of the distance differences between a new point (x) and an existing point (y)

$$D_E = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \quad \text{..... (2.a)}$$

Manhattan distance is the distance between real vectors using the sum of their absolute difference (of a new point (x) and an existing point (y)

$$D_m = \sum_{i=1}^{k}|x_i - y_i| \quad \text{.....(2.b)}$$

Hamming distance is used for categorical variables. If the value of the new point (x) and the value of the existing point (y) are the same, then the distance D = 0, otherwise D = 1

$$D_m = \sum_{i=1}^{k}|x_i - y_i| \quad \text{.....(2.c)}$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 0$$

Where K is defined as a number of points to be considered

This model was used since it does not require a training period and which makes it a faster algorithm, unlike other regression models. With this model, the training dataset is stored and used during real-time prediction. Since our data is huge, this becomes one of the good prediction models. KNN uses data and classifies new data points based on similarity measures [18].

3.3 Least Absolute Shrinkage and Selection Operator (LASSO)

LASSO is the regression analysis technique that reduces the absolute value of the regression coefficients [20]. The objective function that is minimized by the LASSO algorithm can be expressed as;

$$\min_{w} \frac{1}{2n}\|Xw - Y\|_2^2 + \alpha\|w\|_1 \quad \text{.....(3)}$$

where w is the coefficient vector comprising coefficients β associated with features, which are the parameters of the model; X is the feature vector; Y is the target vector; n is the number of depth samples in the training

dataset; and the hyperparameter α is the penalty parameter that balances the importance between the sum of squared errors term and the regularization term, which is the norm of the coefficient vector. This model has been chosen due to the difference in coefficients of predictor variables in our data set. The main advantage of this model on this data set is that it learns the linear relationship and shrinks the regression coefficient towards zero by penalizing the regression model using regularization terms together to ensure the sparsity of the coefficients [21].

3.4 Extreme Gradient Boosting tree

The boosted tree is the ensemble method that constructs more than one decision tree. It is an additive regression model in which individual terms are simple trees. Boosting combines classifiers, which are created from weighted versions of the learning sample, with the weights adaptively adjusted at each step to give increased weight to the cases which were misclassified in the previous step [22]. There are many types of Boosted tree methods including Gradient Boosting, Extreme Gradient Boosting (XGBoost), Stochastic Gradient Boosting and Adaboost. In this study, Extreme Gradient Boosting was applied. Extreme Gradient boosting sequentially adds predictors to the ensemble and follows the sequence of correcting preceding predictors to arrive at an accurate predictor. This model is chosen because it combines the strengths of two algorithms: regression trees (models that relate a response to their predictors by recursive binary splits) and boosting (an adaptive method for combining many simple models to give improved predictive performance) [23].

3.5 Random Forest Regression (RFR)

RFR is another ensemble method that constructs more than one decision. It is a tree-based algorithm whose trees (that are independently trained) are assembled by bagging. According to Hanafy and Mahmoud (2021), a random model for forest regressors can be expressed as follows.

$$g(x) = f_0(x) + f_1(x) + f_2(x) + \ldots\ldots + f_n(x) \quad (4)$$

Where 'g' is the final model that is the sum of all models and each model f(x) is the decision tree. This model was chosen for the given dataset because it combines many decision trees to predict a more accurate outcome. Each tree is used to generate a prediction for a new random sample, then the predictions are averaged to form the forest's prediction [14]. This model reduces the problem of overfitting on our dataset.

4. Estimation of the accuracy of the Prediction

The estimation of the accuracy of the prediction of the above models was evaluated by R squared (R^2), The Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). R^2 is the coefficient of decision. The value of R^2 is between 0 and 1. The higher the R^2, the better the model output. This means the model has deviated less from real values. Thus, the best possible score is 1.0. R^2 is given by

$$R^2 = \frac{Explained\ variance}{Total\ variance} \quad (5.a)$$

The RMSE of the disparity between the expected values and the real values is determined as the square root. The lower the RMSE the better (means are less variance among the expected values and the real values)

$$RMSE = \frac{1}{N}\sum_{n=1}^{N}(\acute{Y} - Y)^2 \quad \text{..... (5.b)}$$

Where N = Number of overall observations, Ý = expected premium value and Y = real insurance premium value

The MAE is the difference between the original and forecast values obtained by averaging the absolute difference over the data set. The lower the MAE the better.

$$MAE = \frac{1}{N}\sum_{n=1}^{N}\left|\acute{Y} - Y\right| \quad \text{..... (5.c)}$$

4. Materials and methods

4.1 Data source

The dataset used for this study was collected from *Kaggle.com* (machine learning repository). The dataset contained medical information and costs billed by the health insurance company. It had 1,339 rows and 7 columns. The following are the columns (variables) in the dataset; age, gender, BMI, number of children, smoking, region and insurance charges. In regression analysis, the value of a dependent variable is predicted using independent variables. While the age, gender, BMI, number of children, smoking and region are treated as independent variables, and insurance charge was an independent variable.

Table 2: Description of the variables in the dataset

Variable	**Description**	**Data type**		**R - Data type (atomic class)**
		Data type	**Categories**	
Age	Age of the primary beneficiary	Continuous		Integer
Sex	Sex of the primary beneficiary/Contractor (male or female)	Categorical (binary)	Male Female	Character
BMI	Body Mass Index, providing an understanding of body weights that are relatively high or low relative to height	Continuous		Numeric
Smoking	Smoking habit of insurance beneficiary (smoking or not)	Categorical (binary)	Yes No	Character
Children	Number of children covered by health insurance, number of dependents	Continuous		Integer
Region	beneficiary's residential area	Categorical	Northeast Southeast Southwest Northwest	Character
Charges/ Expenses	Individual insurance premiums billed by health insurance	Continuous		Numeric

4.2 Data analysis

The data analysis was conducted using R-software. R is powerful data analysis software and has been widely

applied in regression analysis. The relation between predictors or independent variables was explored to test for multicollinearity problem. Whenever an independent variable is highly correlated with one or more of the other independent variables, it can be said that a multicollinearity problem exists [24]. Multicollinearity test found that there was no multicollinearity problem.

The following variables were nominal categorical data; sex, smoker and region. The Linear and KNN models require that all predictor variables be numeric, because categorical data cannot be properly handled by this model. On the other hand, the tree-based models; Random Forest and Gradient boosting model naturally handle numeric or categorical predictors. To get better performance of all the models involved, the categorical data were transformed into numerical data by using a dummy encoding technique which leaves one group out (the first level of the factor) and create new columns for all other groups coded 1 or 0 depending on whether the original variable represented that value or not.

The Predictive modelling was conducted in two phases; training the data set and testing the model. The dataset was split into two parts; the first part was used for model training and the other part was used for model testing. The data set was split into training data and testing data, whereby 80% of the total data was used as training data to train every model used in this study, and then the models were tested using the test data. MLR was used to find the relationship between outcome and associated independent variables. Furthermore, the values of MAE, RMSE, and R2 for each of the models were compared to evaluate the performance of four machine learning algorithms used in the study.

5. Results

5.1 Descriptive statistics

The number of male and female respondents was nearly the same, and more than half of them (79.5%) were non-smokers (Table 5).

Table 1: Descriptive statistics for categorical data - Sex, Smoking and Location variables

Variable		**figure**	**%**
Sex	Female	662	49.5
Male		676	50.5
Smoking	Smoker	274	20.5
Non-smoker		1064	79.5
Location	Northeast	324	24.2
Northwest		325	24.3
Southeast		364	27.2
Southwest		325	24.3

The average age and BMI were 39 years and 30.67 respectively. The average medical charge was $ 13,270 (Table 6).

Table 2: Descriptive statistics for continuous data - Age, BMI, Expenses

	Age	BMI	No. of children	Insurance charges ($)
Minimum	18	16	0	1,122
1st Qu	27	26.28	0	4,734
Median	39	30.4	1	9,382
Mean	39	30.67	1	13,270
3rdQu	51	34.7	2	16,687
Maximum	64	53	5	63,770

The charges vary greatly from around $1,000 to $64,000. Many respondents were charged Less than $2,000 by health insurance companies.

5.2 Multicollinearity test

The Pearson t-test was conducted to find out if the correlation among the predictors (independent variables) was significant. The test results showed no significant correlation among the independent variables. Therefore, it was concluded that the issue of multicollinearity in the dataset did not exist.

5.3 Relationship between insurance charges and predictor variables

This section presents the relationships between insurance charges and the predictor variables used in the study. The fist presentation is relationship between insurance charges and sex. The second one is the relationship between insurance charges and number of children. The third one is the insurance charges and smoking habit. Other presentations included the relationship between health insurance and Age and BMI.

The distribution of insurance charges on sex shows that there is no difference between males and females. The distribution is nearly the same for both sex categories (figure 1).

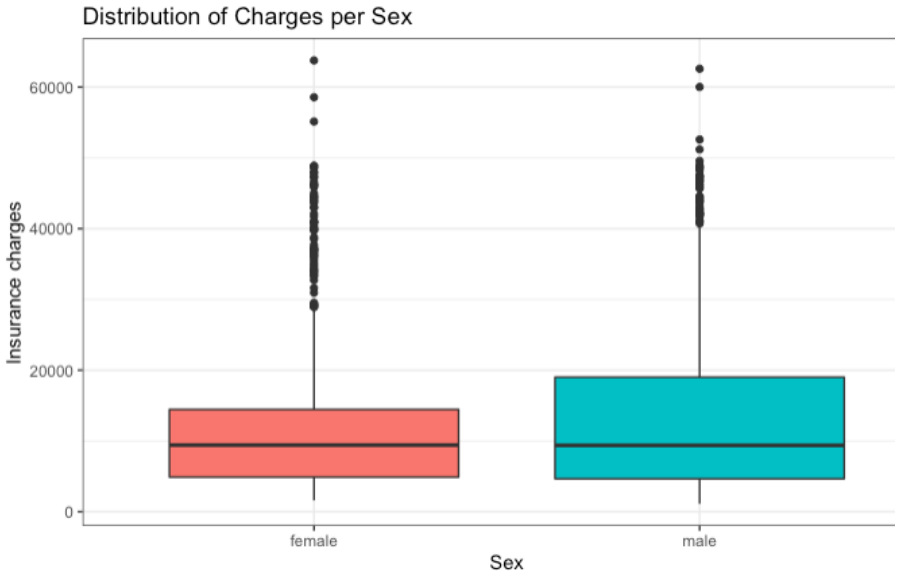

Figure 1: Boxplot of insurance charges per sex

The number of children had a small effect on the insurance charges. The distribution of charges on the number of children shows that most patients that did not have children were charged less than those with children as shown in the diagram below (Figure 2).

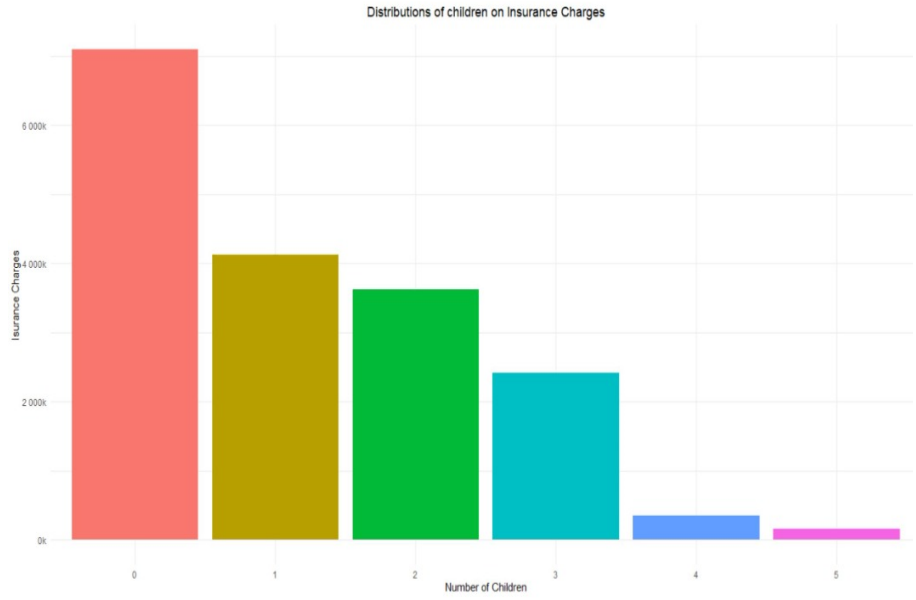

Figure 2: The distribution of insurance charges per children

Regarding charges and smoking, the boxplot (figure 3) indicates that there is a high increase in insurance charges for people who smoke compared with people who do not smoke.

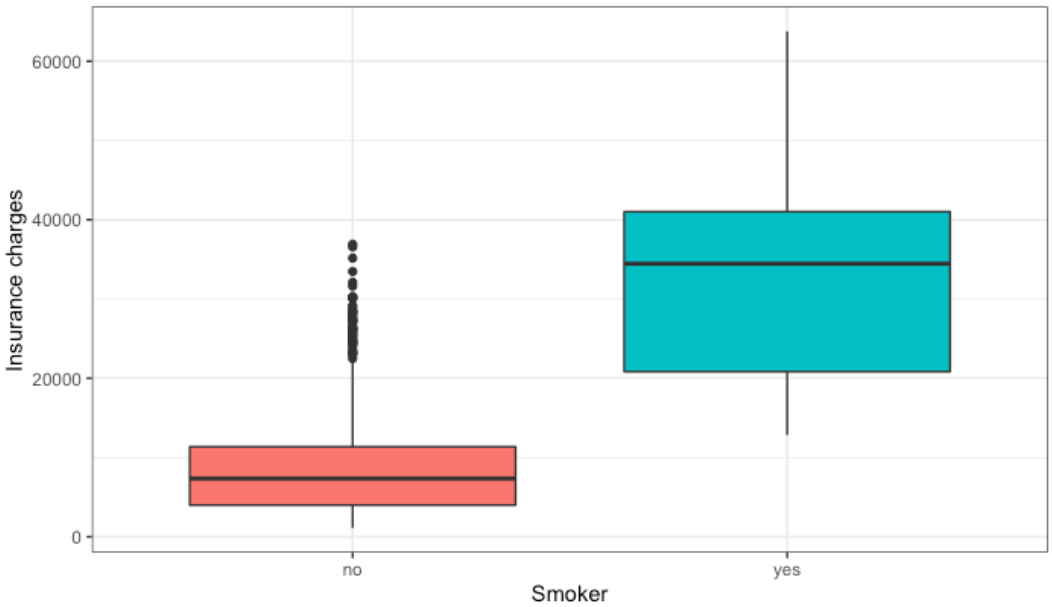

Figure 3: Boxplot of insurance charges per smoking

The plot of age on insurance charges shows that insurance charges increase as age increases (figure 4).

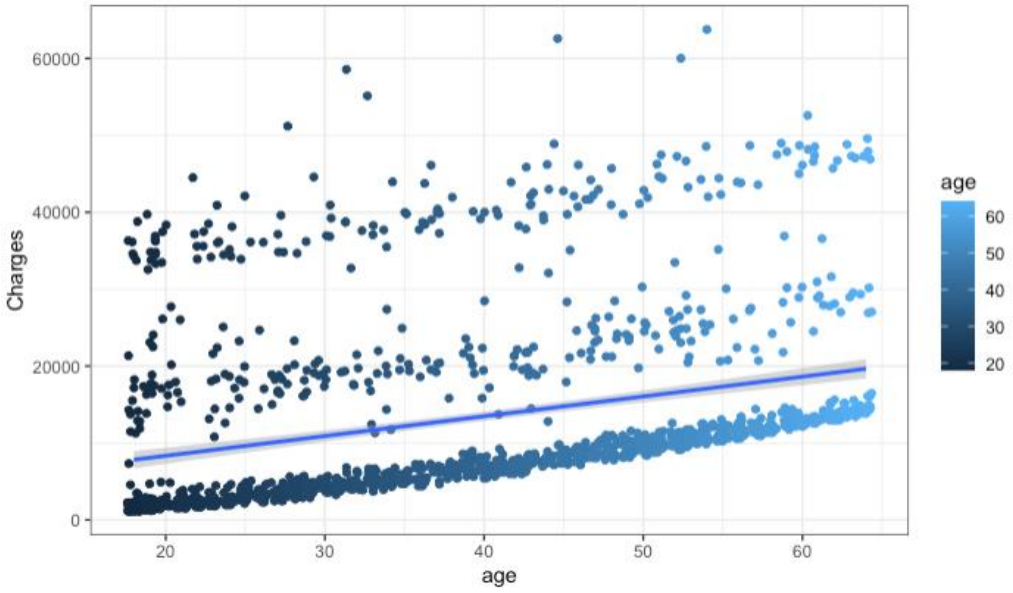

Figure 4: Relationship between insurance charges and age

The plot of BMI on insurance charges shows insurance charges increases as BMI increases (figure 5). This means people with high BMI are charged more by the insurance company than people with low BMI.

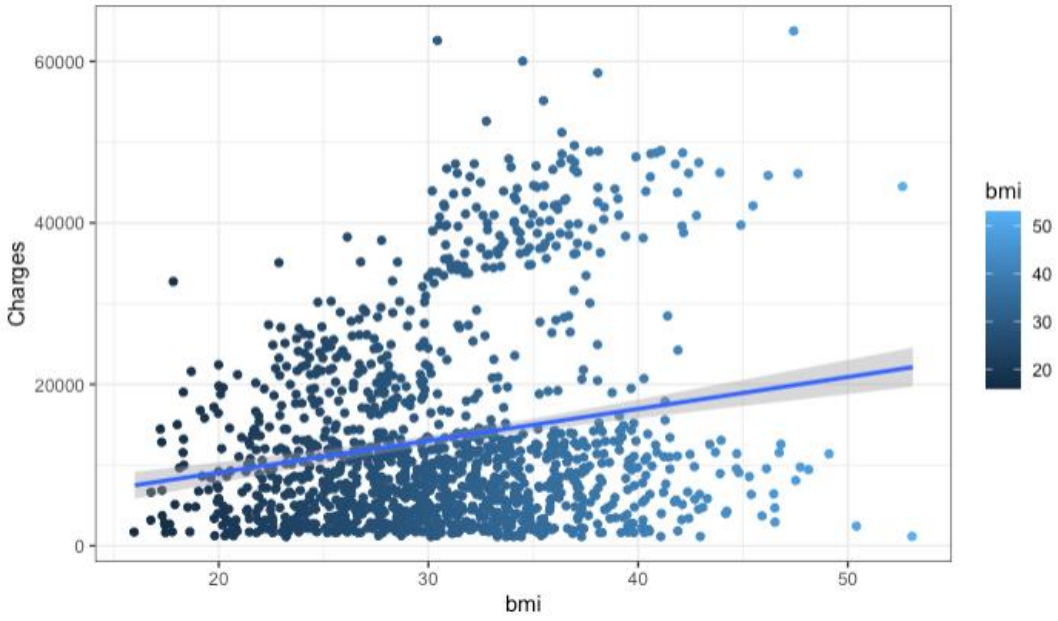

Figure 5: Relationship between insurance charges and BMI

5.5 Predictive Modelling

5.5.1 Multiple liner regression analysis

The most common machine learning regression model used is the MLR. The plot of actual vs predicted values

for MLR show how well the model fits the data (Figure 6). It can be seen that the model performed well in predicting the value for lower charges. However, the prediction performance of the values for higher charges was not well performed

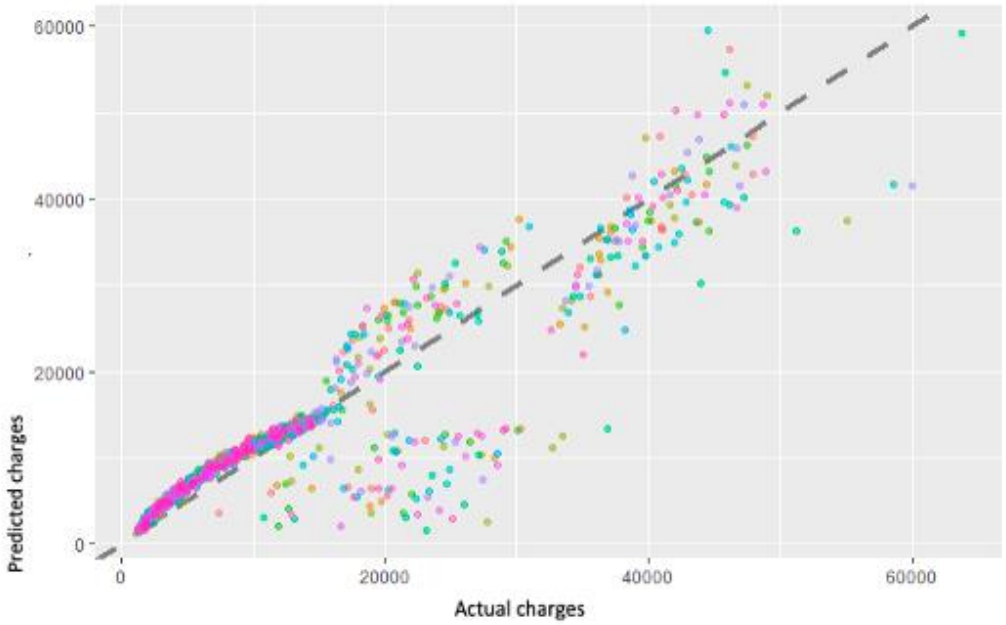

Figure 6: Plot of Actual vs Predicted values for MLR

The coefficients obtained from MLR are tabulated in table 7, and a *p-value* of 0.05 is used as a cut-off point to determine the significance of the variables. The bigger value coefficients represent higher relevance to the model. Based on the results, the following values had a statistically significant influence on the insurance prices; Age, BMI, Smoking and Region. This means that as age and BIM increase, the medical insurance charges increase. Smoking and Region were entered as dummy variables; thus, their interpretation follows the dummy variable interpretation is a little bit different. Regarding Smoking, a significant value means smokers significantly influence insurance charges by one unit when compared with non-smokers. Regarding region, individuals living in the Northeast have higher insurance charges when compared with individuals living in the Southeast region (reference variable)

Table 3: Multiple Linear Regression Variables' coefficients

	estimate	**Std. Error**	**t- value**	**Pr (>I t I)**
(intercept)	3.86147262	0.02992126	129.0544873	0.000***
Age	0.21956219	0.01298423	16.9099069	0.000***
Sex_male	-0.03380324	0.02489352	-1.3579133	0.175
BMI	0.03218319	0.01304476	2.4671357	0.001**
Smoker_yes	0.70958159	0.03130212	22.6688045	0.000***
Children	0.47568900	0.01378000	3.4520000	0.100
Northeast	0.03462494	0.03554844	0.9740211	0.033*
Southwest	-0.03571256	0.03624123	-0.9854124	0.325
Southeast	-0.07255654	0.03629282	-1.9991982	0.046

Sign. Codes: 0 '***' 0.001 '**' 0.01 '*'. 0.05 '-' 0.1 ' ' 1

Based on the estimates and significant values presented in table 8, the MLR predictive model (equation 2) can be presented as follows

$$\text{Insurance Charges} = \beta_0 + \beta_1 \text{Age} + \beta_3 \text{BMI} + \beta_4 \text{smoker} + \beta_5 \text{ Northeast (location)}$$

$$\text{Insurance charges} = \beta_0 + 0.219 \text{ Age} + 0.032 \text{ BMI} + 0.709 \text{ Smoker} + 0.033 \text{ Northeast}$$

5.5.2 Evaluation of the models' performance

To evaluate the performance of four machine learning algorithms, the values of MAE, RMSE, and R^2 for each of the models were compared. MLR has much lower performance on test data and leads to overfitting on this data and would not be preferred for this problem. The findings are tabulated in Tables 9 and 10.

Table 4: Evaluation metrics for train data

Model	R^2	MAE	RMSE
K-nearest Neighbours (KNN)	0.8754706	2691.375	4299.594
Least Absolute Shrinkage and Selection Operator (LASSO)	0.7722119	4136.346	5733.035
Extreme Gradient Boosting	0.8901612	2224.492	3990.497
Random Forest Regression (RFR)	0.8978284	2196.060	3847.263

Table 5: Evaluation metrics for test data (Models' performance comparison)

Model	R^2	MAE	RMSE
K-nearest Neighbours (KNN)	0.8252715	3221.427	5236.708
Least Absolute Shrinkage and Selection Operator (LASSO)	0.7397349	4516.719	6384.349
Extreme Gradient Boosting	0.8553049	2688.200	4748.673
Random Forest Regression (RFR)	0.8530681	2726.356	4783.827

Based on test data, by looking at the value of R^2, the Extreme Gradient Boosting model was able to explain 85.5% of the variation, followed by RFR which explained 85.3%. The RMSE estimate for Extreme Gradient Boosting is 4748.673, which is much better than other models. Based on the findings above, the comparison graph was developed. Figure 8 and 9 presents a comparison of the four (4) models, on three (3) performance measures used in this study (R^2, MAE, RMSE).

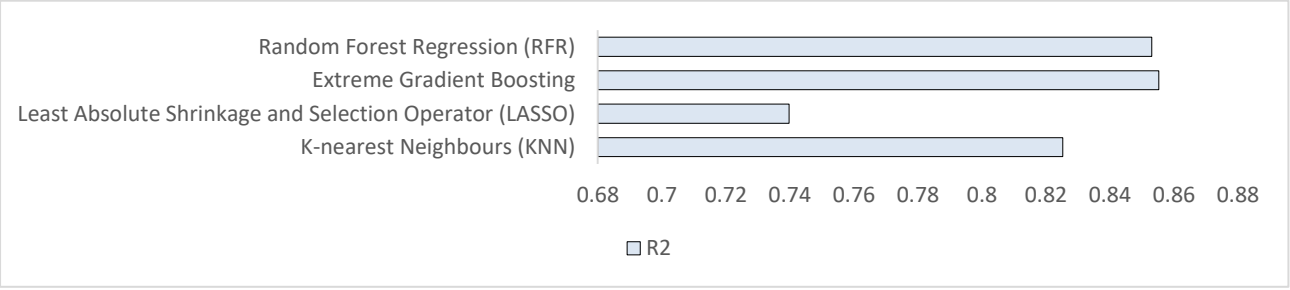

Figure 8: Comparison of the four (4) models, on R^2 performance measure

Figure 9: Comparison of the four (4) models, on two performance measures (RMSE and MAE)

6. Discussion

The number of male and female respondents was nearly the same and the distribution of the data based on the location of the respondents was also nearly the same. More than half of the respondents were non-smokers. The variables that show relationship to medical insurance charges are age, the number of children and smoking habit. Plot distributions of smoking and charges show that there is a strong influence on smoking habit and medical charges. A number of children and Smoking habits were seen to influence the charges. The relationship between Children and insurance charges per smoking shows charges among smokers are higher than charges among non-smokers. However, as the number of children increases, there is a very small to no increase of charges among these groups. Sex is less likely to influence the charges, as the premiums changes had similar distribution for both males and females.

The MLR results show that there is a significant influence of age on medical insurance charges. Therefore, older individuals pay more than younger ones. This might be because old age is associated with diseases including heart diseases and bone-related diseases. Similar findings were also obtained from the studies by [11,13,25,26].

The positive significant relationship between smoking habit and insurance charges means smokers are more likely to be charged higher than non-smokers. This means the insurance company charges higher the smokers than non-smokers. The reasons might be because smoking is associated with diseases such as lung cancer and Tuberculosis. Smoking was also found to be one of the factors that strongly influence health insurance charges by [13,26]. Other previous studies with similar findings are studies by [11,25]

The significant relationship between BMI and Medical charges means people with higher BMI are charged more by insurance companies than people with relatively lower BMI. The reason might be that BMI is often associated with the occurrence of diseases hence this might be the reason for this distribution. Similarly, the study by [11,26] also found a significant influence of BIM on medical insurance charges. Location is another

factor that has a significant influence on medical charges insurance. The medical insurance companies tend to charge differently based on the beneficiary's residential area. These findings are contrary to the findings by [13, 26] which found Region/location has no significant influence on medical charges.

The relationship between the sex of the patient and medical insurance charges was also analysed, whereby sex was found to have no significant influence on medical charges. Thus, there is no or little influence on sex on medical insurance charges. Similarly [26] also found no significant influence of sex on medical charges. Contrary to these findings, the author [12] that found a significant influence of sex on medical insurance charges. Number of children was found to have no significant influence on the medical charges of the beneficiary.

To evaluate the performance of predictive models that use machine learning algorithms to predict health insurance premiums, the performance of five (5) machine learning regression models was evaluated. Those models are MLR, KNN, LASSO, RFR. The MLR had lower performance on test data and lead to overfitting based on the given data, for that reason, this model was not included in the comparison. The performances of the two models; Extreme Gradient Boosting Tree and RFR were better than all other models used. In this regard the Extreme Gradient Boosting Tree and RFR have high ability to appropriately capture linear and non-linear relationships between the dependent and independent variables (as compared to other models evaluated). These findings are supported by other studies that found RFR as the best predictive model when compared with other models [11,14]. Generally, these results give us more reason to recommend the Extreme Gradient Boosting model and RFR in the prediction of health premium rates.

7. Conclusion

Demographic characteristics are the main factors that influence the premium charges among the patients. Age, BMI, smoking, and region play significant roles in predicting the insurance changes. This means the health insurance companies can use these variables to develop a predictive model that might insure fair premium charges to their members. Moreover, the Extreme Gradient Boosting and RFR are recommended as the best model for predicting health insurance premiums. It is expected that premiums predictions based on these models would give realistic payment models that is affordable to many clients as well as enable companies to continue with business. Since the insurance rates that are affordable and give quality services encourage many people to purchase health insurance products [3] usage of these models has a potential to expand universal health coverage

8. Study limitations

The first limitation of this study was lack of primary. The study planned to use primary data from insurance companies in East Africa. However, due to an outbreak of the COVID 19 pandemic, Collection of primary data from these insurance companies which would involve office visitations, face to face meetings and other logistics was not possible. The researcher had to use secondary data. This has limited the researcher an opportunity to use fresh data collected in the East African context. Despite of these fact, it is believed that the findings obtained

from this study will contribute to the efforts to improve the health insurance sub-sector in many countries including the East African countries. Another limitation is the existence of few variables in the dataset. The data set used had only six variables; age, sex, BIM, number of children, smoking, and region. Previous studies found there are additional attributes that have a significant influence on health insurance charges such as education level [14], clinical information and disabled status [15], occupation, type of plan used [17] and past medical history [25]. Future studies can use these attributes to further validate the performance of the ML models.

References

[1] M. Tungu *et al.*, "Does health insurance contribute to improved utilization of health care services for the elderly in rural Tanzania? A cross-sectional study," *Glob. Health Action*, vol. 13, no. 1, 2020, doi: 10.1080/16549716.2020.1841962.

[2] A. Ho, "Health Insurance," *Encycl. Glob. Bioeth.*, 2015, doi: 10.1007/978-3-319-05544-2_222-1.

[3] R. Douven, R. van der Heijden, T. McGuire, and F. Schut, "Premium levels and demand response in health insurance: relative thinking and zero-price effects," *J. Econ. Behav. Organ.*, vol. 180, pp. 903–923, Dec. 2020, doi: 10.1016/j.jebo.2019.02.030.

[4] S. Greenlaw and D. Shapiro, *Principles of Economics 2e*. 2011. [Online]. Available: https://d3bxy9euw4e147.cloudfront.net/oscms-prodcms/media/documents/Economics2e-OP_s2jF42u.pdf

[5] B. Lantz, *Machine Learning with R: Expert techniques for predictive modeling, 3rd Edition*. Packt Publishing, 2019. [Online]. Available: https://books.google.co.tz/books?id=iNuSDwAAQBAJ

[6] WHO, *Consitution of the World Health Organization*, no. October. 2008. doi: 10.4324/9780203029732.

[7] M. Huber *et al.*, "How should we define health?," *BMJ*, vol. 343, no. 7817, 2011, doi: 10.1136/bmj.d4163.

[8] J. F. Outreville, "Theory and Practice of Insurance," *Theory Pract. Insur.*, no. June 2016, 1998, doi: 10.1007/978-1-4615-6187-3.

[9] C. Hong Wang, Kimberly Switlick, Christine Ortiz and and B. Z. Connor, "Africa Health Insurance Hand Book: How to make it work," no. June, 2010, [Online]. Available: www.healthsystems2020.org

[10] C. Rapaport, "An Introduction to Health Insurance : What Should a Consumer Know ?," *Congr. Res. Serv.*, pp. 7–5700, 2015.

[11] A. Lakshmanarao, C. S. Koppireddy, and G. V. Kumar, "Prediction of medical costs using regression algorithms," *J. Inf. Comput. Sci.*, vol. 10, no. 5, pp. 751–757, 2020.

[12] M. hanafy and O. M. A. Mahmoud, "Predict Health Insurance Cost by using Machine Learning and DNN Regression Models," *Int. J. Innov. Technol. Explor. Eng.*, vol. 10, no. 2, pp. 137–143, 2021, doi: 10.35940/ijitee.c8364.0110321.

[13] T. Kaur, "Factors affecting health insurance premiums : Explorative and predictive analysis Factors Affecting Health Insurance Premiums : Explorative and Predictive Analysis Creative Component Project Report By," 2018.

[14] N. Yego, J. Kasozi, and J. Nkrunziza, "A Comparative Analysis of Machine Learning Models for Prediction of Insurance Uptake in Kenya," *MDPI*, no. October, 2020, doi: 10.20944/preprints202010.0186.v1.

[15] C. Yang, C. Delcher, E. Shenkman, and S. Ranka, "Machine learning approaches for predicting high cost high need patient expenditures in health care," *Biomed. Eng. Online*, vol. 17, no. S1, pp. 1–20, 2018, doi: 10.1186/s12938-018-0568-3.

[16] P. Killada, "Data Analytics using Regression Models for Health Insurance Market place Data," University of Toledo, 2017.

[17] P. Strawiński and D. Celińska-Kopczyńska, "Occupational injury risk wage premium," *Saf. Sci.*, vol. 118, pp. 337–344, Oct. 2019, doi: 10.1016/j.ssci.2019.04.041.

[18] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput. Sci.*, vol. 2, no. 3, pp. 1–21, 2021, doi: 10.1007/s42979-021-00592-x.

[19] N. D. Bhadja and P. A. A. Abhangi, "A review Of Machine Learning Methodology in Big data," *Int. J. Sci. Dev. Res. - IJSDR*, vol. 3, no. 5, pp. 361–368, 2018.

[20] Y.-H. Kiang, "Chapter 2.- Model development and validation methodology: A classical big data application," in *Fuel Property Estimation and Combustion Process Characterization*, 2018, pp. 11–39. doi: 10.1016/B978-0-12-813473-3.00002-7.

[21] S. Misra, H. Li, and J. He, "Chapter 5 - Robust geomechanical characterization by analyzing the performance of shallow-learning regression methods using unsupervised clustering methods," in *Machine Learning for Subsurface Characterization*, Elsevier Inc., 2020, pp. 129–155. doi: 10.1016/B978-0-12-817736-5.00005-3.

[22] C. D. Sutton, *Classification and Regression Trees, Bagging, and Boosting*, vol. 24, no. 04. Elsevier Masson SAS, 2005. doi: 10.1016/S0169-7161(04)24011-1.

[23] J. Elith, J. R. Leathwick, and T. Hastie, "A working guide to boosted regression trees," no. Ml, pp. 802–813, 2008, doi: 10.1111/j.1365-2656.2008.01390.x.

[24] M. P. Allen, “Chapter 37 - The problem of multicollinearity,” in *Understanding Regression Analysis*, Boston, MA: Springer, 2007, pp. 176–180. doi: 10.1007/978-0-585-25657-3_37.

[25] E. F. Adebayo, O. A. Uthman, C. S. Wiysonge, E. A. Stern, K. T. Lamont, and J. E. Ataguba, “A systematic review of factors that affect uptake of community-based health insurance in low-income and middle-income countries,” *BMC Health Services Research*, vol. 15, no. 1. BioMed Central Ltd., p. 543, Dec. 2015. doi: 10.1186/s12913-015-1179-3.

[26] A. A. Kodiyan and K. Francis, “Linear regression model for predicting medical expenses based on insurance data,” no. December 2019, 2020, doi: 10.13140/RG.2.2.32478.38722.