UNIVERSITY OF NAIROBI

FACULTY OF SCIENCE AND TECHNOLOGY

DEPARTMENT OF MATHEMATICS

Micro-insurance Pricing with XGBoost Model containing Tweedie GLM

PAULINE MORAA MWEBERI
REG NO .I56/76159/2014

PROJECT SUBMITTED TO THE DEPARTMENT OF MATHEMATICS IN PARTIAL FULFILLMENT FOR A DEGREE OF MASTER OF SCIENCE IN ACTUARIAL

Abstract

Microinsurance is a necessary tool that can be widely used to contribute to alleviating poverty. The low income earners' hard earned wealth or businesses can be protected from the insurable risks. The insurers could consider modelling these products to increase insurance penetration hence increasing the country's GDP.

This paper, I have compared three pricing methods using the same data to draw a conclusion using RMSE as a metric. The results of the project have shown that RMSE of the last model, XGBoost with Tweedie GLM. This approach was taken because of the availability of microinsurance data and the one that is available is heavily zero rated.

Tweedie is the most suited generalized linear model since it is a mixture of Poisson-Gamma distribution. The results are so because of benefits such as regularization; parallel processing;handling missing values ;cross validation and effective pruning. All these have been shown in the main project.

Declaration and Approval

I the undersigned declare that this dissertation is my original work and to the best of my knowledge, it has not been submitted in support of an award of a degree in any other university or institution of learning.

21.11.2022

Signature Date

PAULINE MORAA MWEBERI
Reg No. I56/76159/2014

E-mail: paulmoraa035@gmail.com

In my capacity as a supervisor of the candidate's dissertation, I certify that this dissertation has my approval for submission.

21. 11.2022

Signature Date

Prof. Patrick G. Weke
DEPARTMENT OF MATHEMATICS,
UNIVERSITY OF NAIROBI,
BOX 30197, 00100 NAIROBI, KENYA.
E-mail: pweke@uonbi.ac.ke

Dedication

This study is dedicated to my family and friends who have supported me in various ways in my quest for trans-formative knowledge.
I will add a special thank you to Eld. and Mrs. Moyo that have held my hand and seen that I have completed the graduate program. We did it!

I'm eternally grateful. .

KEYWORDS

Insurer	Insurance Company
Insured	The party that has taken the insurance contract
Premium	The amount an insured pays to the Insurer
Policy	Terms and conditions of the insurance contract
IRA	Insurance regulatory Authority
AKI	Association of Kenya Insurers
GDP	Gross Domestic Product
GLM	Generalized Linear Model
UHC	Universal Health Coverage

Acknowledgements

First and foremost I am grateful to the Almighty God for enabling and guiding me throughout this academic journey since preschool to postgraduate. My sincere appreciation goes to my supervisor Prof.Patrick G. Weke, for his insight, encouragement and assistance in writing this dissertation. His invaluable comments and guidance enabled me to complete this research. I also thank Mr.Peter Mecha who helped me pen my ideas and provided guidance on how to work with different statistical programmes. Finally, I must express my very profound gratitude to my parents Lawrence and Florence, brothers, sister and friends for providing me with continuous encouragement, prayers and unfailing support throughout the entire process of researching, writing this thesis and in my years of study. Thank you.

....

Pauline Moraa Mweberi

Nairobi, 2022.

Contents

1 GENERAL INTRODUCTION

1.1 Study Background

Insurance is a tool designed to protect the insured person from financial loss according to the contract in force. Insurance penetration has been affected because of peoples purchasing power and product design. The insurance penetration in Kenya for the year 2020 was 2.3%. Insurance penetration is the total written premium as a ratio of the country's gross domestic product (GDP). 56.39% of the share is owed to non-life policies and 43.61% a reflection of the life policies Magati[8].

The pricing and distribution of the traditional insurance products has automatically out priced a section of the population who would like to mitigate themselves from the insurable risks but lack the purchasing power. This phenomenon affects the lower-middle and the lower income class citizens.More than one-third or 36% of Kenyans are exposed to numerous risks such as economic shocks, death, illness and loss of property due to accidents, natural disasters and calamities. This high incidence of shocks revealed by Mutia[10] underpins the urgent need for risk management solutions for individuals, households and businesses.

Formal risk management solutions such as insurance can be a potential solution. Yet, the FinAccess survey also found that only two percent of Kenyans used insurance as a solution to deal with shocks. These findings highlight the huge gap between risk protection needs and insurance outreach in Kenya. So, what is the reason for this huge gap? Digging deeper, there is a huge disparity in insurance uptake among different income segments. While there is reasonably high usage among the highest income segment (53 percent), it is quite low among the lower-middle and middle income segments (16 per- cent and 28 percent respectively). These two income segments that can be identified as emerging consumers (earning between Sh 20,000 to Sh55,000 per month) constitute the highest proportion of Kenya's population.

However, this segment has stayed away from insurance and opts for traditional, inadequate risk mitigation mechanisms such as social networks to meet spiraling medical, school and funeral costs through harambees. But such traditional community ties are fast weakening amid the urban bustle and many are wishing they had secured an insurance payout, rather than relying on the goodwill of others. This means that a huge chunk of Kenyans remain unprotected against major shocks, while the insurance industry also stays underdeveloped. Focusing on these emerging consumer groups is key to unlocking insurance potential as they have growing incomes and constitute the highest proportion of the population. But first, there is a need for public policy interventions to stimulate the insurance sector. Among other initiatives, this can include conducive regulations and government spending to insure vulnerable groups.

Then, and most importantly, the insurance industry itself has to expand its horizons by not just focusing on a limited population of high, higher - middle income individuals. For insurers, the sheer number of potential customers in the low -income bracket makes this an attractive market. This is where micro-insurance comes on board. The low-income segment can be served well by offering products whose coverage is lower in value than the usual insurance plan and offers considerably smaller premiums that are attractive to the bottom end of the market. This needs a shift in insurers' mindset and a deliberate application of technology and product innovation to offer attractive value propositions.

Unlocking market potential requires targeting uninsured people, particularly the low-income households. The scarcity of data and even insufficient actuarial skills has led to a challenge in the pricing of micro-insurance products. How can insurers ensure better pricing in the micro-insurance market? This is the main focus of this research. Reducing the price barriers to financial products such as this will greatly benefit the vulnerable groups.

1.2 Problem statement

One of the key features of a micro-insurance product is the price. The low-income earners have limited disposable income; they live from hand to mouth. For insurance companies to tap into this uninsured market it needs to price the product appropriately: not too high since it will not be taken up by customers; not too low that the company becomes insolvent after taking up the risk.

Generalized Linear Modelling is a popular pricing model yet it has its limitations: unable to incorporate variables with non-linear trends; unable to effectively model using zero inflated data. Due to the shortcoming of using the GLM our study will employ a more sophisticated model; Tweedie distribution in the XGboost Model to eliminate these shortcomings.

1.3 Study Objective

1.3.1 Main objective

The main objective of the study is to apply Tweedie distribution in the XGboost model in pricing a micro-insurance product.

1.3.2 Specific Objectives

The following are specific objectives

1. To employ a Generalized Linear Model to the data set that has both linear and non-linear parameters.
2. To apply a Generalized Linear Model with Tweedie distribution to the data set that has both linear and non-linear parameters.
3. To test the same data set on the Tweedie distribution with the XGboost model.
4. To compare the goodness of fit by using the RMSE for the three models and compare the outputs.

1.4 Significance of study

The study is importance since it will increase the efficiency, by use of a single step model, and accuracy of risk premium computation in Micro-insurance which is a new trend in the market. Sufficiency in computing the risk premium will in the long run cushion the risk pool therefore the insurer will be able to conveniently meet its obligation. This will ensure policyholder's livelihood is sustained and the shareholder's interest are safeguarded.

2 LITERATURE REVIEW

Micro-insurance pricing is not highly researched like other fields.Scholars like Obuba(2014)[11], Platteau(2017)[14], Eling(2014)[7],Brau(2011)[4] Biener(2013)[3], Pena(2019)[13] have discussed different angles of it.These authors have written an excellent overview on various aspects of micro-insurance like Sanchez used some multiple regression in conjuncture with the Tweedie model where the case he has used glm coefficients in his results. Churchill(2002)[5] is interested in role of commercial insurers. In Maleika(2008)[9] suggested to advice Social funds maybe directed to cater for micro-insurance to benefit poor people.

In these papers they have touched on various aspects of micro-insurance which have not been practically efficient. Akotey(2011)[2] explains how the demand for micro-insurance vehicles is in modern markets of developing countries like Ghana. He figured out the various factors that may influence the demand of micro-insurance for the sector markets. The markets of developing countries is entrenched in poverty leading to an untapped opportunity. There is low-level of income in those households and even that little that is available is mismanaged. Due to insufficient insurance knowledge by the people may lead to misunderstanding of the concept and therefore the opportunity to take up those insurance covers is missed.He collected data from four major markets in Accra, Ghana and analyzed it using the probit regression model to determine whether the people in the informal sector would purchase insurance or not. All factors withstanding there is potential demand for micro-insurance service is high.

E. Ohlsson(2010)[12] have extensively expounded on the subject of Non-life Insurance Pricing with Generalized Linear Models as stated. Generalized linear models are very resent models . They are more appropriate for pricing compared to ordinary linear models in two ways. The first is, GLMs uses a variety of distributions like normal, Poisson, Gamma e.t.c considering the characteristics of the data being analysed. This is contrary to the constraint of the Ordinary Linear Models that assumes normality of data being modelled. Secondly, the mean of linear models is a linear function of the explanatory variables while for GLMs some monotone transformation of the mean is a linear function of the $x_s^{'}$ with linear and multiplicative cases. Other advantages as noted by [12] include: it has well established techniques for estimating statistical features like standard errors, confidence intervals and testing model selection: it has software packages that are used to tariff analysis. In our case we will be using R for our analysis.

Gordon (2002)[15] wrote on fitting Insurance claims Data on Tweedie's Compound Poisson Model this was in a bid to produce a fair and accurate tariff based on aggregated claims insurance data. They used the Poisson distribution for the claims count and Gamma distribution to model the claims amount data. They noted that it was important to use double generalized linear model; in their case the number of claims and the claim amount; to enable one to model one dispersion in the absence of the other. In practice, the explanatory variables will have different weightings on the dependent variable. Some variables will impact the claims frequency more while other variables will affect the severity more than it would the frequency. Since modelling the aggregated claims insurance data will likely lead to a display of non-constant dispersion data; modelling the dispersion as well as the mean to obtain an efficient estimation of μ. The dispersion parameter used in the sub- model was ψ being that Tweedie models are scale invariant. The residual maximum likelihood (REML)method was used to estimate the variance estimators though it has a downward biases effect when the number of parameters used to estimate the fitted values is large visa vi the sample size. $\hat{\psi}$ is the maximum likelihood estimator. They concluded that the Tweedie Compound Poisson method is efficient in analysing aggregated claims data which is usually used to generate the pure risk premium. Therefore, more terms are likely to be significant in the model in comparison with approximate methods or methods used based on univariate likelihoods.

Obuba(2014)[11] in his 2014 paper has noted down factors that affect micro-insurance product pricing which are:

- Data Availability
- Business acquisition expenses
- Administration cost
- Random stochasticity of the frequency and severity of claims.
- Lack of a suitable actuarial approach

Micro-insurance is a tool that protects the livelihood of low-income earners. This is a alternative to the traditional insurance products. Features of this kind of product were highlighted by Ahmed(2006) [1] as below:

- Targets low-income earners in the economy.

- Product is usually bundled with other financial products
- The product is simple in nature.
- Short underwriting process with few requirements.
- The Sums Insured (assured) per policy are small.
- Short claims management turnaround time.
- Simple distribution channel

Guha-Khasnobis advised that the micro insurer should consider the pricing of these products with small periodic premiums. Deblon(2011)[6] has given incite on five risks that can be considered for modeling micro- insurance products which are: Harvest failure; Third-party liability; property damage; life-cycle risks and health risks. These risk factors are individual and predictable hence good enough for them to be mitigated by having an insurance pool.

Sanchez(2019) [13] used an innovative approach by mixing two distributions - The compound Poisson distribution and the Gamma distribution to come up with a single risk premium model called Tweedie model. He noted that the properties of the data being used will determine the choice of distributions. The model has evidenced its efficiency by giving different risk premium depending on the group risk profile. This is contrary to the inefficient burning cost method that is often used to give a flat premium not incorporating the lapsed policies with a good risk profile or insureds with bad risk profile. The Tweedie model leaves out the extra step of adding an expense loading which is usually part of the burning cost model. The data used to price the micro-insurance product was from both internal and external sources. This is particularly helpful because it covers for the deficiency that actuaries encounter when modeling the price of a micro-insurance product. He has recommended for the model to be replicated with data from different regions and different micro-insurance products. Therefore, this model can be replicated with data from the Kenyan market.

3 METHODOLOGY

3.1 Introduction

In this chapter, we will present the methods for arriving at the earlier stated project objectives. It looks at the model and model development with a clear presentation of the model assumptions. The chapter also presents the sources of data.

3.2 Data Scope

The developing and introduction of micro-insurance products into the market has led to increase in market share. It has however proven impossible to get granulated data from the respective insurance companies who are sighting data protection policies. Access to data possess as the greatest huddle hence testing the model's effectiveness. One particular difficulty of predicting such claims is that there are many policyholders with no accidents and consequently, no claims. The resulting distribution is therefore positively-skewed, with a mass at 0. Such zero-inflated distributions may not fare well if loss functions such as Gaussian distribution are used. We therefore use secondary data that have the following variables:

- Age
- Gender
- Number of Children
- Smoking status

3.3 Generalized Linear Model(GLM)

Generalized linear models represent the class or family of regression models which models the response variable, Y, and the random error term (ε) based on exponential family of distributions such as normal, Poisson, Gamma, Binomial. GLM assumes that the distribution of the response variable is a member of the exponential family of distribution. This is different from the general linear models (linear regression / ANOVA) where response variable, Y, and the random error term (ε)have to be based solely on the normal distribution. Linear models can be expressed in terms of expected value of response variable as the following

$$g(\mu) = \beta_0 + \sum_{i=1} \beta_i X_i \tag{1}$$

where μ can be expressed as $E(Y)$ but for generalized linear models $g(\mu)$ takes the link function associated to the exponential family distribution i.e for Poisson model

$$\log(Y) = \beta_0 + \sum_{i=1} \beta_i X_i \tag{2}$$

3.3.1 component of Generalized linear model

GLM has three components namely

- **Random Component** - specifies the probability distribution of the response variable; e.g., normal distribution for Y in the classical regression model, or binomial distribution for Y in the binary logistic regression model. This is the only random component in the model; there is not a separate error term.
- **Systematic Component** - specifies the explanatory variables $(x_1, x_2, \cdots, x_k)$ in the model, more specifically, their linear combination; e.g.,$\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$, as we have seen in a linear regression, and as we will see in the logistic regression in this lesson.
- **Link Function** $g(\mu)$, specifies the link between the random and the systematic components. It indicates how the expected value of the response relates to the linear combination of explanatory variables; e.g, $g(\mu)$=$g(E(Y_i))$=$E(Y_i)$ for classical regression, or $g(\mu)$=log(μ) for Poisson regression.

3.3.2 Assumptions of GLM

The following are assumptions of the generalized linear models

- The data $Y_1, Y_2, \cdots, Y_n$ are independently distributed.
- The dependent variable Y_i does NOT need to be normally distributed, but it typically assumes a distribution from an exponential family (e.g. binomial, Poisson, multinomial, normal, etc.).

- A GLM does NOT assume a linear relationship between the response variable and the explanatory variables, but it does assume a linear relationship between the transformed expected response in terms of the link function and the explanatory variables.
- Explanatory variables can be nonlinear transformations of some original variables.
- The homogeneity of variance does NOT need to be satisfied. In fact, it is not even possible in many cases given the model structure.
- Errors need to be independent but NOT normally distributed.
- Parameter estimation uses maximum likelihood estimation (MLE) rather than ordinary least squares (OLS).

3.4 Tweedie distribution

The Tweedie distribution is named after M. C. K. Tweedie 1984. Tweedie studied the exponential dispersion model with power variance functions. This kind of distribution can have a cluster of data items at zero. This property makes Tweedie distribution so useful in modelling any zero-inflated data sets such as micro-insurance. It is a mixture of a a Poisson and Gamma distribution.

Let N; be the number of claims;it follows a Poisson distribution with mean λ ; it is bound between 0 to ∞.

Let Y_i be the claim amount that is independent and identically distributed. Data-set will generate a Gamma distribution with parameters α and γ.

$$S_N = Y_1 + Y_2 + Y3 + Y4 + Y_5 + \cdots + Y_N \tag{3}$$

The variable S_N is a Poisson sum of Gamma. The probability of not having a claim is:

$$P(S_N = 0) = e^{-\lambda}$$

The p.d.f of the Tweedie distribution for $y > 0$ is therefore

$$f_s(y) = \sum_{n=1}^{\infty} e^{-\lambda} \frac{\lambda^n}{n!} \frac{\gamma^{n\alpha}}{\Gamma(n\alpha)} y^{n\alpha-1} e^{-y\gamma} \tag{4}$$

The mean and variance will be:

$$E(S_N) = \lambda \frac{\alpha}{\gamma} \tag{5}$$

and

$$Var(S_N) = \frac{\lambda\alpha}{\gamma^2}(1+\alpha) \tag{6}$$

Changing the Tweedie parameters λ, α, $\frac{1}{\gamma}$ to Poisson-Gamma parameters we get:

$$\lambda = \frac{\mu^{2-p}}{\phi[2-p]} \tag{7}$$

$$\alpha = \frac{2-p}{p-1} \tag{8}$$

$$\frac{1}{\gamma} = \phi(p-1)\mu^{p-1} \tag{9}$$

Substituting Tweedie's new parameters above to equations (2) will be;

$$f_s(y) = exp\left[\frac{-1}{\phi}\left[\frac{\mu^{2-p}}{2-p} + \frac{y}{[p-1]\mu^{[}p-1]}\right] + S(y,\phi)\right] \tag{10}$$

If $1 < p < 2$, then $Y > 0$, with a positive mass at $Y = 0$; refer to the histogram to make this observation.

Note that if:

- $p = 0$ it becomes a normal distribution
- $p = 1$ it becomes a Poisson distribution
- $p = 2$ it becomes a Gamma distribution
- $P = 3$ it becomes an inverse Gaussian Distribution

Note that $1 < p < 2$ We will apply this model on the data in the next chapter.

3.5 XGBoost Model

This is a tree-based algorithm. It is an improvement from the initial decision tree; to Bagging; to random forest; Boosting; gradient boosting; then to the popular XGBoost. Decision tree showed a graphical representation of the of problem solutions under set conditions then conclusions are drawn. The development to Bagging which groups together meta-algorithms that combines predictions from multiple-decision trees. Then came another improvement; random forest where only a subset of features are selected at random to build a collection of decision trees. The came Boosting technique that reduces the error from previous models while increasing performance models. A later improvement to Gradient Boosting technique led to deployment of a gradient descent algorithm to minimise errors as the model progresses. Then came a later improvement to the XGBoost which we will apply in the project to solve a risk premium prediction.

The running of the model will lead to getting the best parameters that will best fit the explanatory variables. The XGBoost has the following features that make it a viable option for actuaries to use when pricing the micro-insurance products. These features are:

- Prevents over-fitting
- Accommodates missing values in your data-set
- Can be run on large data - It is a black box
- Can cross-validate at each iteration
- Auto-pruning is possible

We will derive the XGBoost model before we apply it to our data.

$$Obj(\theta) = l(\theta) + \Omega(\theta) \tag{11}$$

$Obj(\theta)$-Objective function.

$l(\theta$-Loss function.

$\Omega(\theta)$-Regularization function.

The loss function in equation () measures how predictive the model is in relation with the data being modelled.

$l(\theta) = \sum_i (y_i - \hat{y}_i)^2$

The regularization function $\Omega(\theta)$ in equation () controls the complexity of the model while performing iterations to avoid over-fitting.

$$Obj = \sum_{i=1}^{n} l\,(y_i - \hat{y}_i)^t + \sum_{i=1}^{t} \omega(f_i)$$

The objective function at step t will be:

$$Obj(t) = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \omega(f_t) \tag{12}$$

where;

$$\tag{13}$$

$$g_i = \delta_{\hat{y}_i{}^{(t-1)}} l\left(y_i, \hat{y}_i{}^{(t-1)}\right) \tag{14}$$

$$h_i = \delta^2_{\hat{y}_i{}^{(t-1)}} l\left(y_i, \hat{y}_i{}^{(t-1)}\right) \tag{15}$$

$\omega(f)-$ is the complexity of the tree.
Note that the function of the tree is:

$$f_t(x) = \omega_{q(x)}, \omega \in R^T, \quad q: \quad R^d \rightarrow [1, 2, 3, \cdots, T] \tag{16}$$

$\omega-$ vector of scores on leaves.
$q-$ Function assigning each data point to its corresponding leaf.
$T-$ Number of leaves.
Therefore:

$$\omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} \omega_j^2 \tag{17}$$

Substituting equation () into equation () we will arrive at;

$$\begin{aligned} Obj^(t) &= \sum_{i=1}^{n}[g_i\omega_{q(x_i)} + \frac{1}{2}h_i\omega^2_{q(x_i)}] + \gamma T + \frac{1}{2}\lambda\sum_{i=1}^{T}\omega_j^2 \\ &= \sum_{j=1}^{T}\left[\left(\sum_{i\in I_j} g_i\right)\omega_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right)\right]\omega_j^2 + \gamma T \end{aligned} \tag{18}$$

Where $I_j = (i/q(x_i) = j)$ - Are the set of indices of data points assigned to the j^{th} leaf.

$$G_j = \sum_{i\in I_j} g_i H_j \qquad\qquad = \sum_{i\in I_j} h_i$$

Hence, the objective function will be:

$$Obj^(t) = \sum_{j=1}^{T}[G_i\omega_j + \frac{1}{2}(h_I + \lambda)\omega_j^2] + \gamma T \tag{19}$$

Pruning is a very useful tool of the model since it prevents over-fitting.
Pruning of over-fitted data is done when Gain<λ i.e the split will not occur.

$$Gain = \frac{1}{2}\left[\frac{G_l^2}{H_l+\lambda} + \frac{G_r^2}{H_r+\lambda} - \frac{(G_l+G_r)^2}{H_l+H_r+\lambda}\right] - \gamma \tag{20}$$

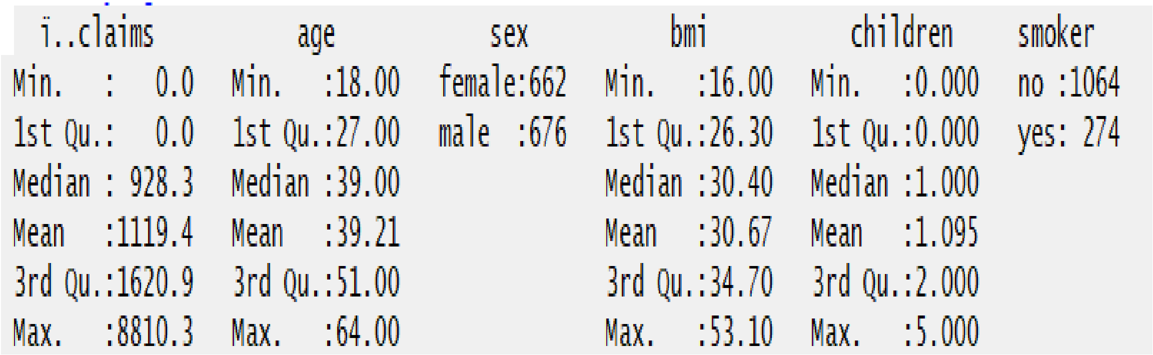
We will therefore go to the next chapter to implement this model.

4 DATA ANALYTIC AND INTERPRETATION

4.1 Introduction

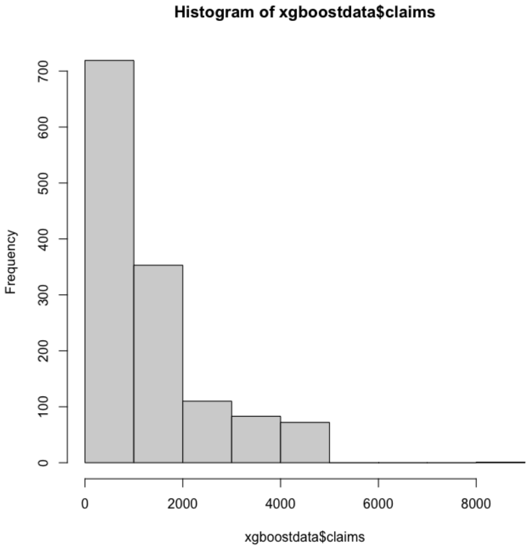
In this section we will compare the two pricing models with the same data set. The Data was run with RStudio programming language. The codes are attached for reference.

Figure 4.1,shows the characteristics of the response and explanatory variables that have been used to run the two models.



```
   ï..claims          age            sex           bmi           children        smoker
 Min.   :   0.0   Min.   :18.00   female:662   Min.   :16.00   Min.   :0.000   no :1064
 1st Qu.:   0.0   1st Qu.:27.00   male  :676   1st Qu.:26.30   1st Qu.:0.000   yes: 274
 Median : 928.3   Median :39.00                Median :30.40   Median :1.000
 Mean   :1119.4   Mean   :39.21                Mean   :30.67   Mean   :1.095
 3rd Qu.:1620.9   3rd Qu.:51.00                3rd Qu.:34.70   3rd Qu.:2.000
 Max.   :8810.3   Max.   :64.00                Max.   :53.10   Max.   :5.000
```

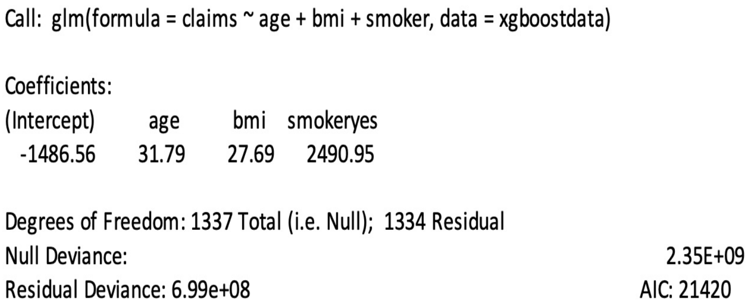
A look at the below histogram figure



The data contains 1338 entries has more that a quarter of it with no claim reported; the histogram has captured the same visually. The age range of the policyholders is between 28 years and 64yrs having an almost equal proportion in the gender disparity. The Body Mass Index goes from a minimum of 16 to 53.10. The policyholders have between 1 to 5 children recognised in the policy details. 3% of them are smokers. The independent variables presented contribute to the frequency and severity of claims being brought forth. We will use three methods to investigate the same data and later infer the model with the best results. The three models are: Simple Generalised Linear model; Generalised Linear Model with Tweedie distribution; Tweedie distribution with xgboost model. We will interrogate one at a time

4.2 Generalized Linear Model

A GLM was run on the same data that produced the below result.

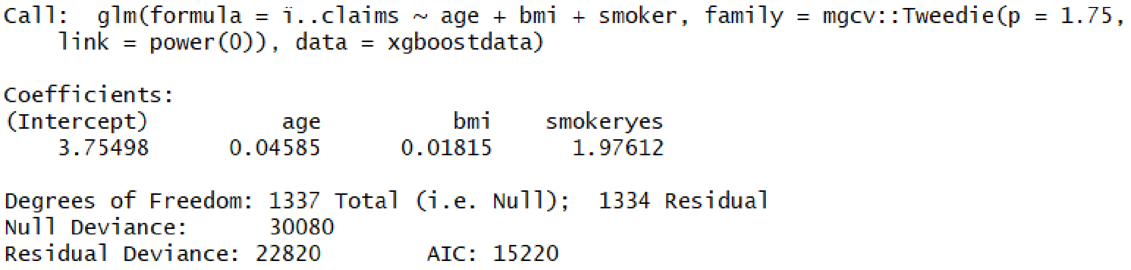


```
Call:  glm(formula = claims ~ age + bmi + smoker, data = xgboostdata)

Coefficients:
(Intercept)        age        bmi   smokeryes
  -1486.56       31.79      27.69     2490.95

Degrees of Freedom: 1337 Total (i.e. Null);  1334 Residual
Null Deviance:                                                  2.35E+09
Residual Deviance: 6.99e+08                                     AIC: 21420
```

The model runs on assumptions like:linear relations between the dependent and the independent variables; the independent variables don't have multicollinearity correlation; the error components are normally distributed which might not be the case for all the data sets. Before the model is fitted; all these aspects need to be investigated. Therefore, scouting for a more robust model.

The Beta of the smoking status has the highest number concluding that this explanatory variable has the highest relation with the claim amount. The least contributing factor is Body Mass Index. The age of the policyholder has the second highest weighting. The RMSE of this regression model is 722.8002. Its AIC equally on the higher side.

4.3 Generalized linear model Tweedie distribution



```
Call:  glm(formula = ï..claims ~ age + bmi + smoker, family = mgcv::Tweedie(p = 1.75, 
    link = power(0)), data = xgboostdata)

Coefficients:
(Intercept)          age          bmi    smokeryes  
    3.75498      0.04585      0.01815      1.97612  

Degrees of Freedom: 1337 Total (i.e. Null);  1334 Residual
Null Deviance:      30080 
Residual Deviance: 22820        AIC: 15220
```

RMSE for Generalised Linear Model with the Tweedie distribution is 641.591. Its AIC is 15220. This model runs the Generalised Linear Model including the Tweedie distribution because of the unique characteristics of the micro insurance data.

4.4 Tweedie distribution with xgboost model

The data that is being used to run this model must be changed to sparse metrics. Gender and smoking status are the variables that were affected since they were factor variables.

The ability of an actuary to change the parameters while modeling is a significant step towards finding the best fit model. This can be done automatically by use of the function $Grid-search$ We did change the parameters eta and max_depth and the output difference can be seen in the figures from(fig:4) to figure(fig:6)below.

$nrounds$ defines the maximum number of iterations/trees.

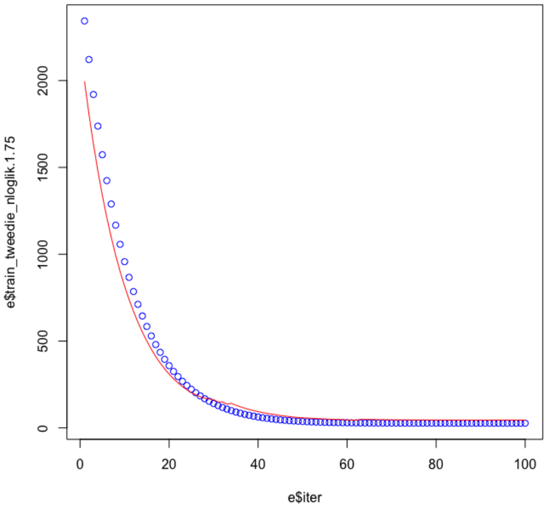
max_depth instructs the number of branched the tree can have; increases the chance of the model to view interactions. It is worth noting that the a higher value increases the complexity of the model. The recommended range is $1-5$ but it can take values $1 \to \infty$.

eta- learning rate; defines the gradient of the errors. The lower the value of eta the higher the $nrounds$. The recommended range is between 0 and 0.3.

gamma is set at zero to prevent from over-fitting the model. *var.power* while running insurance date ranges between [1,2].

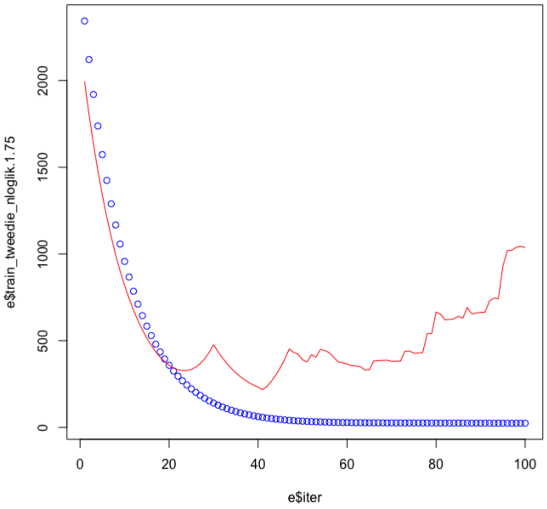
4.5 Sensitivity analysis.

XGBoost comes with a built-in cross validation tool for each iteration. However, to see the effect of different model parameters I have a run sensitivity analysis as below figure 1 : eta = 0.1 and the max depth= 3



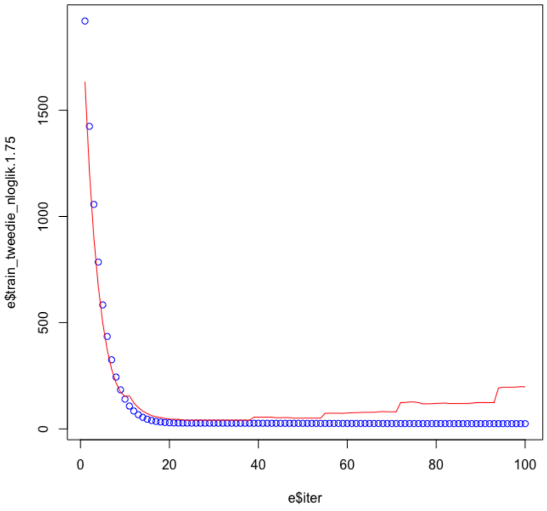
The graph generated with these parameters have lead to under-fitting of the train data compared to the test data all through the hundred iterations.

figure 2 : eta = 0.1 and the max depth= 5



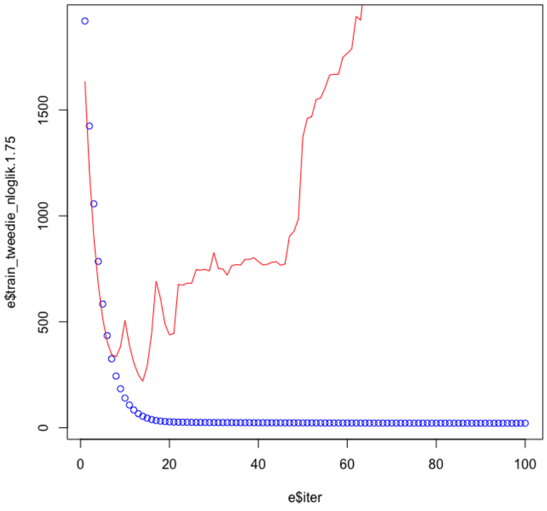
The trend noted with the inputed parameters indicate a sudden variation of the standard error hence not the best parameters to set on the model.

figure 3 : eta = 0.3 and the max depth= 3



It is the best model parameters. The standard error between the train data and the test data is minimal even as the number of rounds increases. The RMSE of the train data was optimal at round 62 at a value of 478.7296. This makes the approximation of the claim amount considering the respective explanatory variables reliable. It will neither be overfitted nor under-fitted.

figure 4 : eta = 0.3 and the max depth= 5



The graph shows that there was a great deal of over fitting. With these parameters of output has the most drastic standard error therefore becoming unfit to be used to simulate the claims amount.

4.6 Models comparison

MODEL	RMSE
GLM	722.8002
Tweedie GLM	641.591
Tweedie GLM in XGBoost Model	478.7296

It is evident that the RMSE of the three models have been improving as we went along.Therefore the Xgboost with Tweedie distribution is the best model since it has the lowest RMSE.

5 CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

The Simple GLM model doesn't account for correlation between the independent variables and non-linear terms. This characteristic therefore takes away from the model's efficiency.

Tweedie distribution with Xgboost model is a robust model choice due to its ability to easily tweak the different model until the desired results are achieved.

5.2 Recommendation

By studying the topic I have come across limitations like; scanty academic research in this area and data inadequacies. This is to encourage others to study this area using available big data from the social sciences since getting primary data from the industry has proven impossible due to data protection. Xgboost model has capacity to a manipulate big data.

The government is strongly pushing the Universal Health Coverage agenda; this can be treated a micro-insurance after population segmenting.This can be an areas of future research. IRA is also offering incubation for start-ups with innovative insurance solutions that would increase penetration.

Insurers should consider making strategic partnerships that will help it with the distribution of the product hence increasing its sales at the minimum cost possible.

A researcher can carry out Sensitivity analysis on λ, tweedie variance power.

Bibliography

[1] M Ahmed. Market research on microinsurance demand. *Technical Assistance Consultant's Report-Market Research on Microinsurance Demand. Project*, (4761-SRI), 2006.

[2] Oscar Joseph Akotey, Kofi A Osei, and Albert Gemegah. The demand for micro insurance in ghana. *The Journal of Risk Finance*, 2011.

[3] Christian Biener. Pricing in microinsurance markets. *World Development*, 41:132–144, 2013.

[4] James C Brau, Craig Merrill, and Kim B Staking. Insurance theory and challenges facing the development of microinsurance markets. *Journal of Developmental Entrepreneurship*, 16(04):411–440, 2011.

[5] Craig Churchill. Trying to understand the demand for microinsurance. *Journal of International Development*, 14(3):381, 2002.

[6] Yvonne Deblon and Markus Loewe. The potential of microinsurance for social protection. *Craig Churchill/Michal Matul (2012)(eds.): Protecting the Poor: A Microinsurance Compendium*, 2, 2011.

[7] Martin Eling, Shailee Pradhan, and Joan T Schmit. The determinants of microinsurance demand. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 39(2):224–263, 2014.

[8] Steve KB Magati. *Influence of distribution channels on consumer purchase behavior in the life insurance industry: a case of Britam Life Assurance Company Kenya.* PhD thesis, Strathmore University, 2021.

[9] Marc Maleika and Anne T Kuriakose. Microinsurance: extending pro-poor risk management through the social fund platform. 2008.

[10] Cecilia K Mutia. *Access to Credit and Household Savings in Kenya Evidence From Kenya National Finaccess 2019 Survey.* PhD thesis, University of Nairobi, 2020.

[11] Chache Willys Obuba. *The effect of product pricing on the growth of micro insurance by insurance underwriters in Kenya.* PhD thesis, University of Nairobi, 2014.

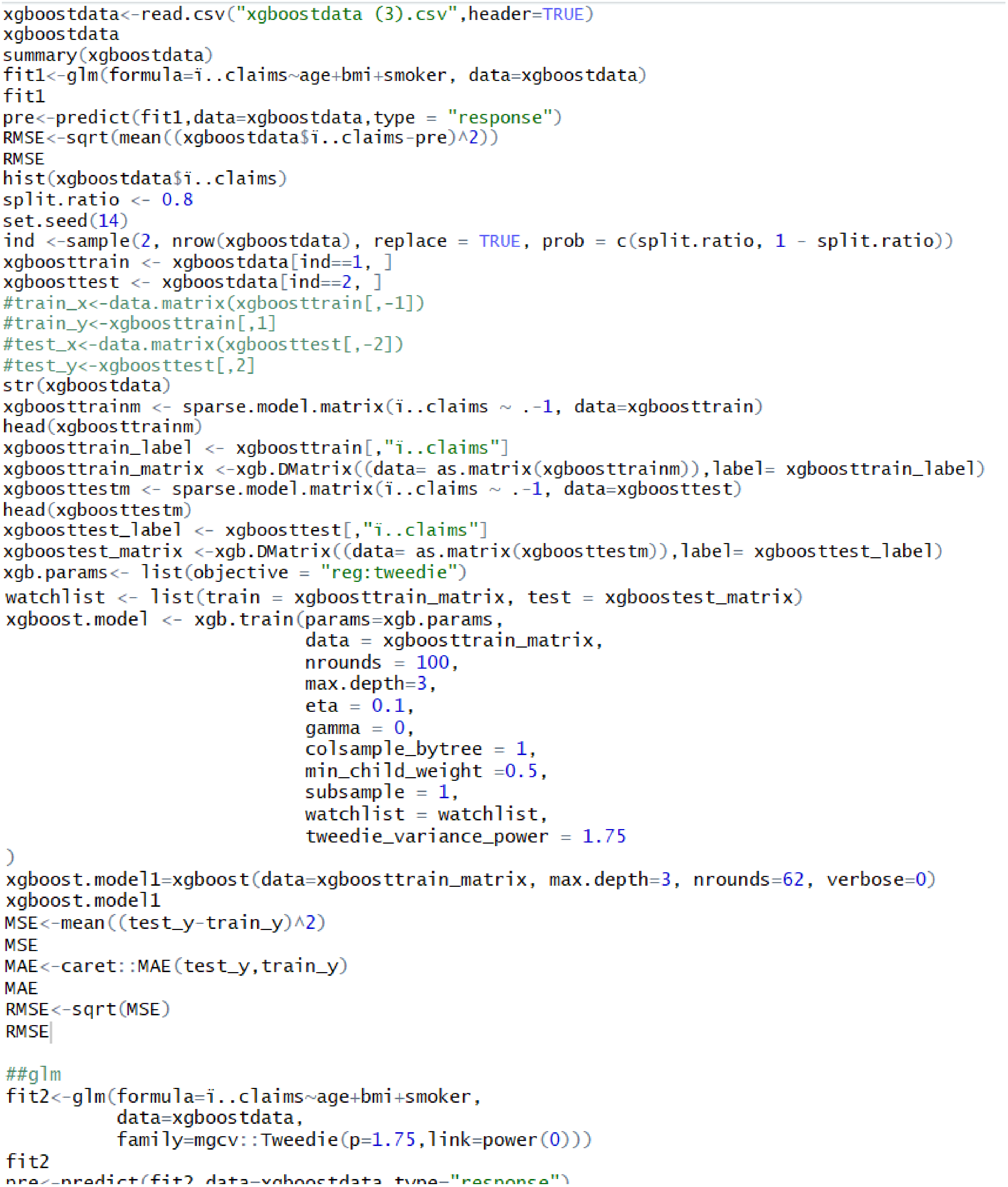
[12] Esbjörn Ohlsson and Björn Johansson. *Non-life insurance pricing with generalized linear models*, volume 2. Springer, 2010.

[13] Inmaculada Peña-Sanchez. Applying the tweedie model for improved microinsurance pricing. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 44(3):365–381, 2019.

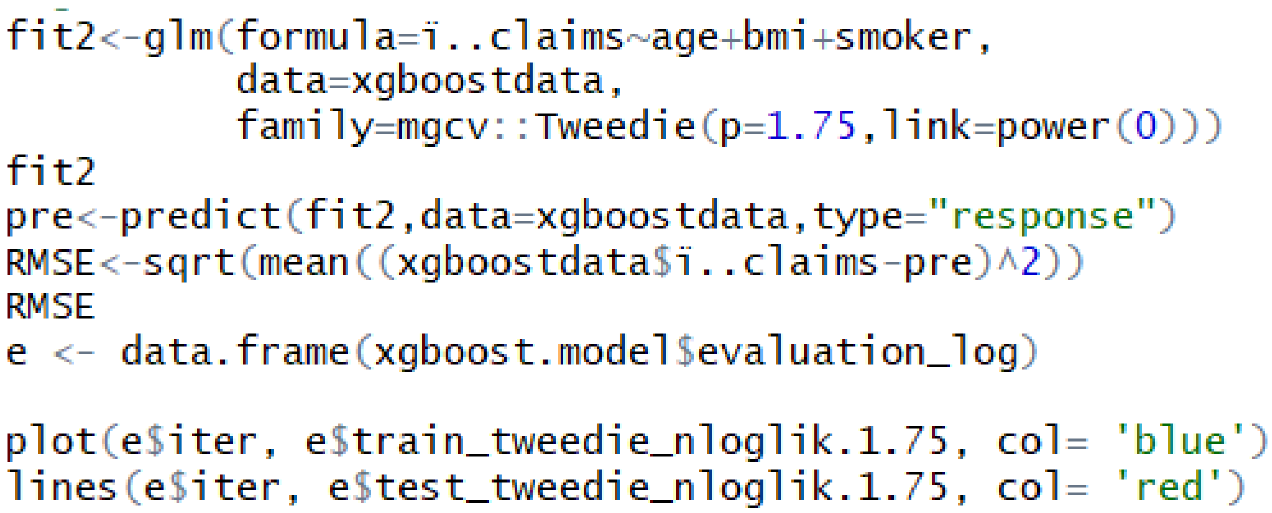
[14] Jean-Philippe Platteau, Ombeline De Bock, and Wouter Gelade. The demand for microinsurance: A literature review. *World Development*, 94:139–156, 2017.

[15] Gordon K Smyth and Bent Jørgensen. Fitting tweedie's compound poisson model to insurance claims data: dispersion modelling. *ASTIN Bulletin: The Journal of the IAA*, 32(1):143–157, 2002.

Appendix



```r
xgboostdata<-read.csv("xgboostdata (3).csv",header=TRUE)
xgboostdata
summary(xgboostdata)
fit1<-glm(formula=ï..claims~age+bmi+smoker, data=xgboostdata)
fit1
pre<-predict(fit1,data=xgboostdata,type = "response")
RMSE<-sqrt(mean((xgboostdata$ï..claims-pre)^2))
RMSE
hist(xgboostdata$ï..claims)
split.ratio <- 0.8
set.seed(14)
ind <-sample(2, nrow(xgboostdata), replace = TRUE, prob = c(split.ratio, 1 - split.ratio))
xgboosttrain <- xgboostdata[ind==1, ]
xgboosttest <- xgboostdata[ind==2, ]
#train_x<-data.matrix(xgboosttrain[,-1])
#train_y<-xgboosttrain[,1]
#test_x<-data.matrix(xgboosttest[,-2])
#test_y<-xgboosttest[,2]
str(xgboostdata)
xgboosttrainm <- sparse.model.matrix(ï..claims ~ .-1, data=xgboosttrain)
head(xgboosttrainm)
xgboosttrain_label <- xgboosttrain[,"ï..claims"]
xgboosttrain_matrix <-xgb.DMatrix((data= as.matrix(xgboosttrainm)),label= xgboosttrain_label)
xgboosttestm <- sparse.model.matrix(ï..claims ~ .-1, data=xgboosttest)
head(xgboosttestm)
xgboosttest_label <- xgboosttest[,"ï..claims"]
xgboostest_matrix <-xgb.DMatrix((data= as.matrix(xgboosttestm)),label= xgboosttest_label)
xgb.params<- list(objective = "reg:tweedie")
watchlist <- list(train = xgboosttrain_matrix, test = xgboostest_matrix)
xgboost.model <- xgb.train(params=xgb.params,
                           data = xgboosttrain_matrix,
                           nrounds = 100,
                           max.depth=3,
                           eta = 0.1,
                           gamma = 0,
                           colsample_bytree = 1,
                           min_child_weight =0.5,
                           subsample = 1,
                           watchlist = watchlist,
                           tweedie_variance_power = 1.75
)
xgboost.model1=xgboost(data=xgboosttrain_matrix, max.depth=3, nrounds=62, verbose=0)
xgboost.model1
MSE<-mean((test_y-train_y)^2)
MSE
MAE<-caret::MAE(test_y,train_y)
MAE
RMSE<-sqrt(MSE)
RMSE

##glm
fit2<-glm(formula=ï..claims~age+bmi+smoker,
          data=xgboostdata,
          family=mgcv::Tweedie(p=1.75,link=power(0)))
fit2
pre<-predict(fit2,data=xgboostdata,type="response")
```



```
fit2<-glm(formula=ï..claims~age+bmi+smoker,
          data=xgboostdata,
          family=mgcv::Tweedie(p=1.75,link=power(0)))
fit2
pre<-predict(fit2,data=xgboostdata,type="response")
RMSE<-sqrt(mean((xgboostdata$ï..claims-pre)^2))
RMSE
e <- data.frame(xgboost.model$evaluation_log)

plot(e$iter, e$train_tweedie_nloglik.1.75, col= 'blue')
lines(e$iter, e$test_tweedie_nloglik.1.75, col= 'red')
```