Journal of Scientific Research & Reports

21(5): 1-10, 2018; Article no.JSRR.31117
ISSN: 2320-0227

Application of Lasso Regression to Model National Development Indicators and National Internet Usage

Yusuf Musa[1*], Babalola Rotimi[2] and Ogedebe Peter[3]

[1]Department of Computer Science, Bingham University Karu, Nigeria.
[2]Andela, Lagos, Nigeria.
[3]Department of Computer Science, Baze University, Abuja, Nigeria.

Authors' contributions

This work was carried out in collaboration between the authors. The authors read and approved the final manuscript.

Article Information

DOI: 10.9734/JSRR/2018/31117
Editor(s):
(1) Dr. Janusz Brzdek, Department of Mathematics, Pedagogical University, Poland.
Reviewers:
(1) Vilda Purutçuoğlu, Middle East Technical University, Turkey.
(2) Radosław Jedynak, Kazimierz Pulaski University of Technology and Humanities, Poland.
(3) P. Wijekoon, University of Peradeniya, Sri Lanka.
Complete Peer review History: http://www.sciencedomain.org/review-history/28136

Original Research Article

Received 21 December 2016
Accepted 03 July 2018
Published 04 January 2019

ABSTRACT

Aim: This paper aim to use Lasso Regression Model to ascertain how the level of development in a country affects the interest of a number of internet users.
Methodology: Least Absolute Shrinkage and Selection Operator (Lasso) regression with the Least Angle Regression selection (LARs) algorithm with k=5-fold cross validation was used to estimate the lasso regression model used to ascertain the significant association between the number of internet user in a country and the development indicators for that country. The change in the cross validation average (mean) squared error at each step was used to identify the best subset of the predictor variables. The lasso regression model was estimated on a training data set consisting of observations from the year 2012 (N=199), and a test data set included the observations from the year 2013 (N=196).
Results: LASSO regression model was trained on N=199 countries and used to identify the best subset of predictors which predicted the response variables; Number of internet users in N=196 countries around the world for the year 2013. The Number of internet users for training and test sets per 100 people for the countries ranged from 1.06 to 96.2 and 1.30 to 96.55 respectively. This indicates that there is significant variation in the response variable.

**Corresponding author: E-mail: yusufmusa@binghamuni.edu.ng;*

Conclusion: It is possible that the few variable indicators we considered as strong predictors of internet are confounded by other factors not considered in the analysis. Therefore, it is recommended that future efforts should focus on other ways to fill in the missing observations since there are large number of national development indicators/factors that are associated with the number of internet users.

Keywords: Regression model; LASSO regression algorithm; LAR algorithm; data mining; model selection.

1. INTRODUCTION

LASSO is an acronym for Least Absolute Shrinkage and Selection Operation proposed by [1]. The Shrinkage is the capability of adding constraints on parameters that shrinks coefficients towards zero, while Selection is used to identifying the most important variables associated with the response variable [2,3]. LASSO is a type of penalization regression method often used in machine learning for linear regression models. LASSO regression methods uses supervised learning methods by penalizing the sum of the absolute values of the weight found in the model, where the sum of a specific penalty is an upper bound [4,5]. This penalty causes regression coefficient of some variables to shrink towards zero and sometimes may even eliminate some of these variables. The shrinkage process allows for better interpretation of the model and identifies the variables that most strongly associates with the target corresponding variables. The process of identifying the most strongly associated variables is referred to as Variable Selection Process. The selected variables are therefore the subset of the predictors that minimizes the predictor error [6].

In this study, LASSO was used over other regularization methods to identify the most significant key factors (that is among a total of 80 variables) that are significantly associated with the number of internet users in a country. To achieve this using LASSO, a turning parameter called Lambda was applied to the regression model to control the strength of the penalty. Such that, where Lambda value increases, more regression coefficients are reduced to zeros, minimizing the number of selected predictors which results to more shrinkage of the non-zero coefficients. Where Lambda is zero, the result is an Ordinary Least Square (OLS) regression analysis [7]. One of the advantages of LASSO can be observed in situations where a large number of observations and the true relationship between the response variables and the predictors are approximately linear.

The aim of this study is to use LASSO Regression Models to determine if there is a significant association between the number of internet users in a country and the development indicators for that country such as GDP per capital, Access to electricity, urban population, mobile cellular subscriptions and the adjusted net national income. The research tends to answer the question: how the level of development in a country affects the interest of a number of internet users in that country. The motivation of the study was to answer the aforementioned research question which relates to almost every country, but it is of great importance to developing countries where a significant number of people do not have reliable internet access. However, it will be remarkable to identify the key factors that are significantly associated with the number of internet users in a country. This will help the cause of connecting more people in developing countries to the internet thereby offering them the opportunities in areas like e-learning in education, e-Payment in civil services, job creation and social networking. These opportunities will help them gain skills, create jobs and provide services which will better not just their lives but the lives of people around them.

2. MATERIALS AND METHODS

LASSO is a model selection method for linear regression which is used when the aim is to select a subset of predictors from a set of initial predictor as defined by the research question in this study. This is achieved by shrinking some predictors to zero. This technique helps to increase prediction accuracy and model interpretability. The regularization capacity of LASSO has made it more useful for feature selection and preventing over-fitting of training data [8,9]. It uses the L1 normalization as a penalty to the sum of absolute value of weight

found to be the regression. The LASSO estimator can be used to achieve the least absolute shrinkage and sparsity using two formulation; firstly, by minimizing the regression error $(y_i - \sum_{j=0}^{m} w_j x_{ij})^2$ subject to $\lambda \sum_{j=0}^{m} |w_j|$. Secondly, by minimizing the regression error $(y_i - \sum_{j=0}^{m} w_j x_{ij})^2$ subject to the error $\sum_{j=0}^{m} |w_j| \leq \eta$. Where $\eta \geq 0$ is the tuning parameter. It is worth knowing that the two equations are the same for any given $\lambda \epsilon (0, \infty)$, there exists $\eta \geq 0$ such that the two problems have the same solution as presented by G. James et al in [10]. In this study the second approach was used as shown in equation (2).

The LASSO estimator ($f(w)$) minimizes the RSS with respect to the measure of the magnitude of the coefficient described in equation (1)

$f(w)$= *RSS (**w**) + λ ** (measure of magnitude of coefficients) (1)

Equation (1) can also be defined as

$$f(x) = \underset{w_j}{minimizes}: \left\{ \sum_{t=1}^{n} \left(y_i - w_0 - \sum_{j=0}^{m} w_j x_{ij} \right)^2 \right\}$$

$$subject\ to\ \lambda \sum_{j=0}^{m} w_j \leq \eta \qquad (2)$$

Where $\lambda \epsilon (0, \infty)$ is a tuning parameter. Lambda is imposed as penalty on the regression sum of squares (RSS) to shrink the size of estimated weight w_j such that as λ increases, more weight w_j are reduced to zero since it tends to shrink the weight w_j by a fixed amount, while a decrease in λ leaves us with OLS regression as stated in section 1.

The choice of tuning parameter λ is usually determined using cross validation in a given training data (x_i, y_i) (such that $i = 1, \dots n,$). In this study N = 395. We then construct an estimator $\hat{f}(w)$ of some unknown function $f(w)$. Suppose that $\hat{f}(w) = f(w)_\lambda$ depends on the tuning parameter λ. To choose a value of turning parameter λ that best optimizes the predictive accuracy of $\hat{f}(w)_\lambda$, cross validation was used. With cross validation the training data was randomly divided into fixed *k-folds(say k=5)*, for each value of the tuning parameter λ, each fold is held out one at a time, while we train on the remaining data and predict the held out observation as presented in equation (4) as also in [11,12].

$$CV(\lambda) = \frac{1}{n} \sum_{k=0}^{k} \sum_{i \epsilon F_k}^{n} (y_i - \hat{f}_\lambda^{-k}(x_i))^2 \qquad (3)$$

We then choose the value of λ that best minimizes our error curve as shown in Fig. 5. In this case the λ were derived using equation (5) with 5-folds cross validations

$$\hat{\lambda} = \underset{\lambda \epsilon \{\lambda_1, \dots \lambda_m\}}{argmin}\ CV(\lambda) \qquad (4)$$

Next, the estimator was inserted into equation (2) to the entire training set (x_i, y_i) such that $i = 1, \dots n,$ using $\hat{\lambda}$. Since this may not give a perfect curve, we have to compute the standard error for the curve. So, for each value we choose as k-fold (that is k=1,… K) we compute the standard error using equation (6) for each n_k point in the kth fold.

$$CV_k(\lambda) = \frac{1}{n_k} \sum_{i \epsilon F_k}^{n} (y_i - \hat{f}_\lambda^{-k}(x_i))^2 \qquad (5)$$

Next, we compute the simple standard deviation for $CV_1(\lambda), CV_2(\lambda), \dots, CV_k(\lambda)$ using equation (6).

$$SD(\lambda) = \sqrt{var(CV_1(\lambda), CV_2(\lambda), \dots, CV_k(\lambda))} \qquad (6)$$

And finally, we compute the standard error using equation (6) as $SE = SD(\lambda)/\sqrt{K}$. The result of change in the validation mean square error at each step is presented in Fig. 5.

2.1 Data Sample

The data sample used was collected from the World Bank dataset of over 80 variables for 248 countries for the year 2012 and 2013 (N = 248). It presents the most current and accurate global development data available, which includes national, regional and global estimates. To perfectly answer the research questions without being biased, all rows of the explanatory variables of the dataset were used.

2.2 Data Measures

The response variable serves as the number of internet users in a country, which is measured by taking the sum total of every 100 persons with internet access in a country. The number of user(s) or person(s) per 100 population is usually used as a unit of measurement of

national indicator. In the case of this study, it is used for measuring the number of internet usage per 100 persons in a nation as presented in [13]. The five predictors used include the following:

1) Percentage of a country's population with access to electricity
2) Number of mobile cellular subscriptions (per 100 persons) in a country
3) Percentage of a country's population living in urban areas i.e. Urban population
4) A Country's GDP per capita (in US$) and
5) A Country's Adjusted Net National Income Per Capita (in US$).

2.3 Data Management

The data management decision taken was to drop missing observations of the explanatory variables from the training and test data sets. After dropping missing data, N = 199 observations of the explanatory variables were left to work with.

3. EXPERIMENTAL SETUP AND DATA ANALYSIS

The distributions for the predictors and the response variables (number of internet users) were evaluated by calculating the mean, standard deviation, minimum and maximum values because they are all quantitative variables. A scatterplot of each individual predictor and the response variable was examined and the Pearson correlation coefficient was used to test the significance of the bivariate association between individual predictors and the response variables. LASSO regression with the Least Angle Regression Selection Algorithm was used to identify the subset of the explanatory variables that best predicts the number of internet users. The LASSO regression model was estimated on a training data set consisting of observations for the year 2012 (N=199), and test data set observations for the year 2013 (N=196). All predictor variables were standardized to have a mean = 0 and standard deviation = 1 prior to conducting the LASSO regression analysis. Cross validation was performed using 5-folds cross validation as shown in equation (5). The change in the cross validation mean squared error rate at each step was used to identify the best subset of predictor variables as presented in Fig. 5. Predictive accuracy was assessed by determining the mean squared error rate of the training data for the prediction algorithm when applied to observations in the test data set as presented in Fig. 4. Python 2.7, a general purpose programming language running on Jupyter notebook 4.2.1. was used in writing the codes. The high performance multidimensional array in Numpy library was used for array computations while the matplotlib library was used for plotting the graphs.

4. RESULTS ANALYSIS

This section presents results of the experiments starting with the descriptive statistics of the predictor and response variable, followed by the bivariate analysis and discussion of estimation and validation of LASSO regression model used in the study.

4.1 Descriptive Statistics

Table 1 shows a summary of the descriptive statistics of the predictors and response variable - Number of internet users.

The average number of internet users (per 100 people) was 38.04 persons per 100 people (sd = 27.47), with a minimum of 1.07 persons per 100 people and a maximum of 96.21 persons per 100 people.

4.2 Bivariate Analysis

The scatterplots were used to visualize the linear relationship between predictors and number of internet users as shown in Fig. 1.

Based on the relationships presented in Fig. 1, Pearson correlation coefficient was used to examine the strength and significance of the association between the response variables and the predictors [14]. Below are the Pearson correlation values (and the p-values) for each predictor and the response variable – Number of Internet Users.

Using Table 2, we can conclude that Access to Electricity and Adjusted Net National Income has the strongest association (**) with the response variable while Mobile Cellular Subscriptions has the weakest association (*) with the response variables among the five predictors.

Table 1. Descriptive statistics of the predictors and response variable

Predictors	N	Mean	Std. Dev	Minimum	Maximum
Access to Electricity (% of Population)	199	77.51	30.09	6.40	100.00
Mobile Cellular Subscriptions (per 100 people)	199	97.98	36.62	7.06	187.36
Urban Population (% of Total)	199	55.44	21.99	8.80	98.95
GDP Per Capita (in US$)	199	13432.70	19036.21	244.20	105447.09
Adjusted Net National Income Per Capita (in US$)	199	10364.14	14329.57	140.28	78441.34
Number of Internet Users (per 100 people)	199	38.04	27.47	1.07	96.21

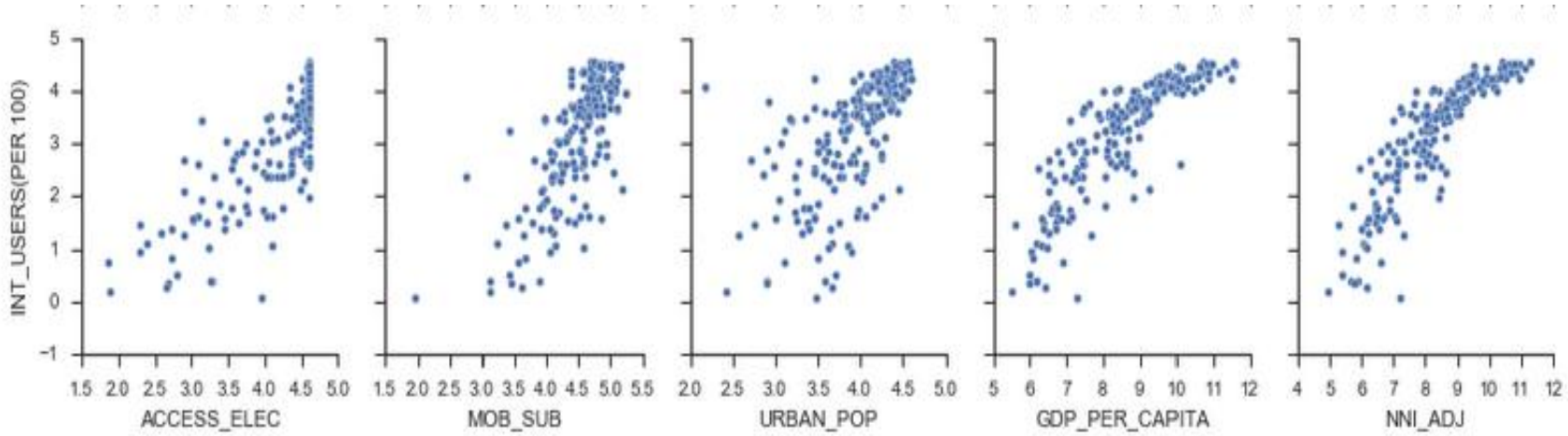

Fig. 1. Association between predictors and number of internet users

Table 2. Pearson correlation for each predictors and response value

Predictors	Pearson correlation (r)	p-value
Access to Electricity	r = 0.81**	P<0.0001
Urban Population	r = 0.73	P<0.0001
Mobile Cellular Subscriptions	r = 0.65*	P<0.0001
GDP per capita	r = 0.77	P<0.0001
Adjusted Net National Income	r = 0.80**	P<0.0001

Since all the variables in use in the analysis are quantitative, scatterplot was also used to examine and visualize the relationship between each predictor and the response variable. In order to aid visualization of the data-points, all the variables were transformed by taking the log before plotting them. This helped to put all the variables in relatively the same scale so that they could be easily compared. Figs. 2 & 3 below show some of the scatterplots.

The scatterplots in Figs. 2 & 3 confirm the positive association between the response variable and the predictors plotted. This means, for example, that internet usage is positively associated with Adjusted Net National Income (NNI) i.e. the number of internet users in a country tends to increase with the NNI of that country.

4.3 Estimation and Validation of Lasso Regression Model

LASSO regression with the Least Angle Regression Selection Algorithm with 5-folds cross validation was used to estimate the LASSO regression model on the training set, which was then validated on the test set [11]. The cross validation was achieved using equation (4) and (5).

The changes in the cross validation average (mean) squared error at each step was used to identify the best subset of predictor variables. However, after running the LASSO regression algorithm, all the predictors remained in the model. Equation 5 and 8 were used to calculate the standard error, this could mean that all the predictors were strongly correlated with the response variable and therefore remained in the model. Figs. 4 & 5 shows the regression coefficient progression for LASSO path and the change in the validation mean squared error at each step respectively.

Considering Table 3, of all the predictors, Adjusted Net National Income (NNI) and Access to Electricity were the most strongly associated with the Number of Internet Users, followed by Number of Mobile Cellular Subscriptions, Urban Population and GDP per capita which had a negative coefficient.

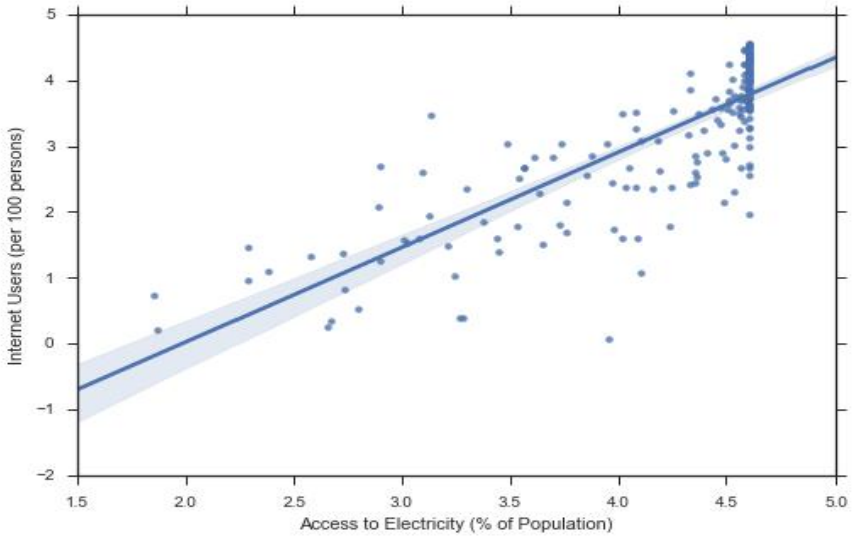

Fig. 2. Scatterplot of Access to Electricity against the number of Internet users

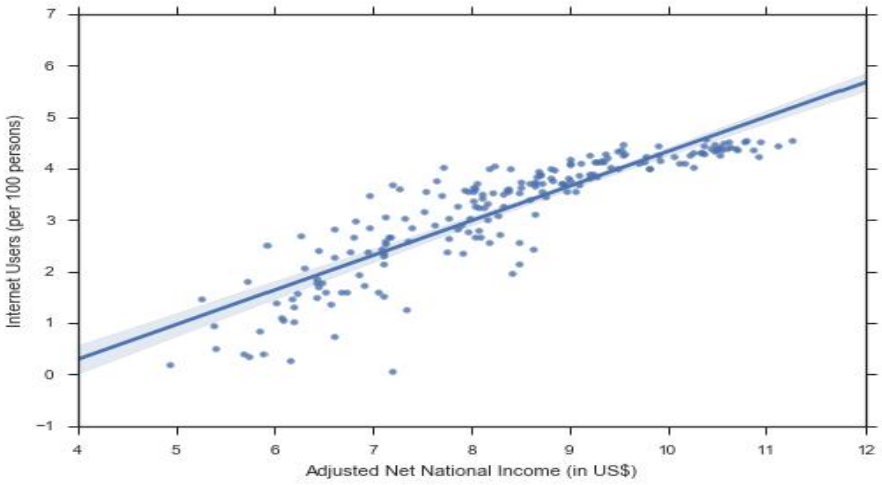

Fig. 3. Scatterplot of Adjusted Net National Income against the number of Internet users

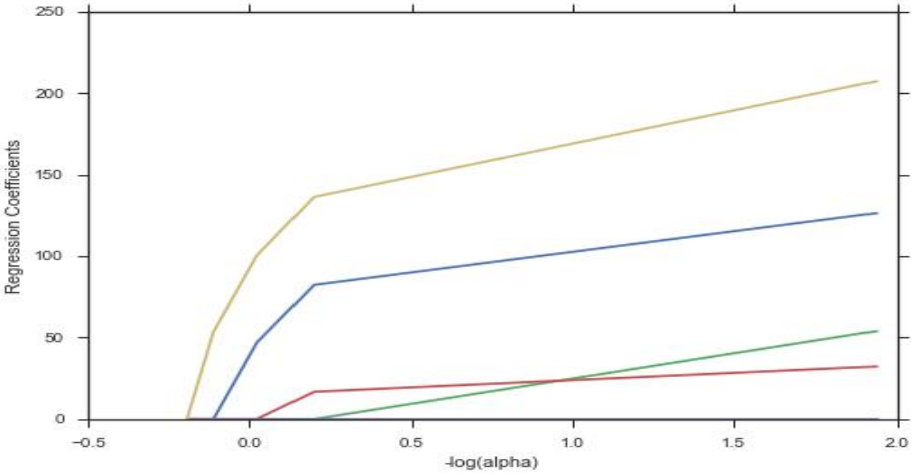

Fig. 4. Regression coefficients progression for lasso paths

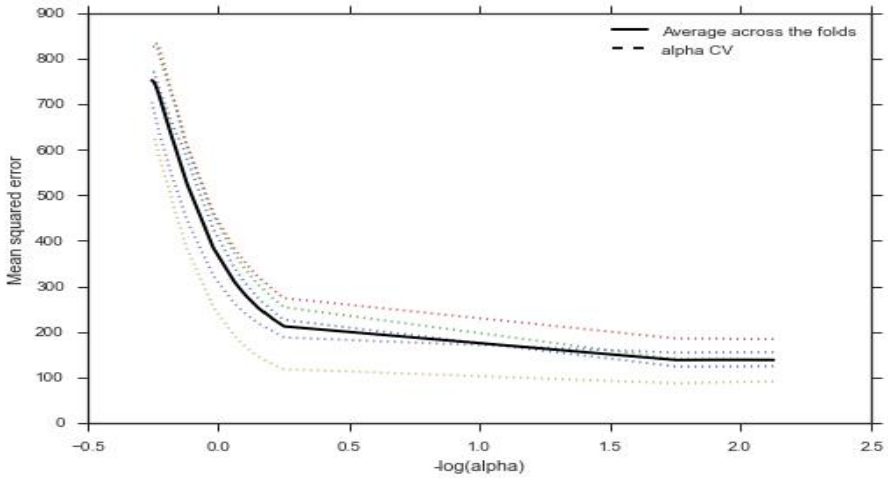

Fig. 5. Change in the validation mean square error at each step

Table 3. The final value of the coefficients for each predictor

Predictor	Coefficient
Access to Electricity	8.762
GDP per capita	-8.783
Mobile cellular subscriptions	4.189
Adjusted Net National Income	23.719
Urban population	2.242

This means that the model will predict a higher number of internet users for countries with high NNI and high access to electricity. Also a high number of mobile cellular subscriptions and high urban population were associated with an increase in the number of internet users. All together the 5 predictors accounted for 83.1% (training data) and 81.6% (test data) of the variation in the number of internet users.

The mean squared error (MSE) obtained for the test data (MSE=143.9) differed slightly from the MSE obtained for the training data (MSE=126.2), which suggests that predictive accuracy did not drop when the LASSO regression algorithm developed on the training data set was used to predict the number of internet users in the test data set.

5. DISCUSSION OF RESULTS AND FINDINGS

In this paper, LASSO regression model (trained on N=199 countries) was used to identify the best subset of predictors that predicted the response variable – Number of internet users in N=196 countries around the world for the year 2013. The Number of internet users for the countries ranged from 1.06 to 96.2 and 1.30 to 96.55 number of internet users per 100 people for the training and test sets respectively. This indicates that there is significant variation in the response variable.

The LASSO regression analysis indicated that all 5 predictors were selected in the final model. These 5 predictors accounted for 81.6% of the observed variability in the number of internet users. The strongest predictors of the response variable were the Adjusted Net National Income (NNI) and the GDP per capita of the country. Furthermore, high access to electricity and high urban population was associated with a high number of internet users.

There was a slight increase in the MSE when the training set LASSO regression algorithm was used to predict the number of internet users in the test set. This suggests that the predictive accuracy of the algorithm may be stable for predicting the number of internet users in the future.

The results of this project indicate that countries with high Adjusted Net National Income (NNI), high access to electricity and high urban population tend to have more internet users. This is plausible because if a county has a high NNI it means that the citizens of that country earn good incomes and can afford an internet connection. Furthermore, if a country has high access to electricity it will have more internet users because all devices that are required for an internet connection e.g. computers, modems, routers, switches etc. are powered by electricity; therefore, if a country has low access to electricity it is unlikely it will have a high number of internet users. Finally, internet access is more common in urban areas than rural areas. This means that countries with low urban population may have a low number of internet users.

In order to increase the number of internet users in developing countries it is suggested that cheap and reliable sources of electricity should be made available in developing countries as suggested by [15,16]. It is important that the energy source be cheap because if it is unaffordable it will not help to increase the number of citizens that have access to electricity. Furthermore, the cost of an internet connection should be made cheaper in low income countries, this will greatly help to increase the number of internet users. Also, the number of internet users in rural areas and communities can be increased by establishing low cost cyber-cafes in these areas so people living there can easily afford and have access to the internet. In addition to this, people in rural areas should be educated on the opportunities the internet offers them so they can utilize its potential to the maximum.

6 LIMITATIONS, RECOMMENDATIONS AND CONCLUSIONS

In this paper we present a predictive model for the number of internet users that appears to have little bias and variance in different samples. In addition, it provides more information on which factors are most likely to have a significant impact on number of internet users. However, there are some limitations that

should be considered when interpreting the results of this project.

Firstly, the predictor 'Access to Electricity' was only available for the year 2012, so there is a need to use the same observations of this predictor in the training and test data which may have affected the predictive accuracy of the model. It would be good to build a model in the future where all observations for this predictor are available for every year.

Secondly, missing data was dropped from the analysis. This has reduced the number of countries that the model was trained on from N=248 to N=199. This meant that the model was left with few data to train on which is bound to affect its out of sample performance. It is recommended that, future efforts should focus on other ways to fill in or input these missing observations.

Finally, there are large number of national development indicators/factors that are associated with the number of internet users, but in this study we only consider a few of these indicators. It is possible that the indicators we considered as strong predictors of internet usage in this study are confounded by other factors not considered in the analysis. As a result, these same factors may not emerge as important predictors when other factors are taken into consideration. Therefore, future efforts to develop a solid predictive algorithm for the number of internet users should expand the algorithm by adding more national development indicators to the statistical model. The plausible selection of lambda via different methods, rather than the cross-validation approach can be considered in future work.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Tibshirani R. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological). 1996;58(1):267–288.
2. Grandvalet Y. Least absolute shrinkage is equivalent to quadratic penalization. In ICANN 98, ed: Springer. 1998;201-206.
3. Oyeyemi GM, Ogunjobi EO, Folorunsho AI. On performance of shrinkage methods–a Monte Carlo Study. International Journal of Statistics and Applications. 2015;5:72-76.
4. Hesterberg T, Choi NH, Meier L, Fraley C. Least angle and ℓ1 penalized regression: A review. Statistics Surveys. 2008;2:61-93.
5. Simon N, Tibshirani R. Standardization and the group lasso penalty. Statistica Sinica. 2012;22:983.
6. Tibshirani R. Regression shrinkage and selection via the lasso: A retrospective. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2011; 73:273-282.
7. Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005;67:91-108.
8. Li J, Das K, Fu G, Li R, Wu R. The Bayesian lasso for genome-wide association studies. Bioinformatics. 2011; 27:516-523.
9. Tibshirani RJ, Taylor J. Degrees of freedom in lasso problems. The Annals of Statistics. 2012;1198-1232.
10. Kim J, Kim Y, Kim Y. A gradient-based optimization algorithm for lasso. Journal of Computational and Graphical Statistics. 2008;17(4):994-1009.
11. James G, et al. An introduction to statistical learning: With applications in R, Springer Texts in Statistics. Springer Science+Business Media New York. 2013;219.
    DOI: 10.1007/978-1-4614-7138-71
12. Tibshirani RJ, Taylor JE, Candes EJ, Hastie T. The solution path of the generalized lasso: Stanford University; 2011.
13. United Nation. Indicators of Sustainable Development: Guidelines and Methodologies – Third edition, page 309. (Access on 13 July, 2018)
    Available:http://www.un.org/esa/sustdev/natlinfo/indicators/methodology_sheets.pdf
14. Hauke J, Kossowski T. Comparison of values of Pearson's and Spearman's correlation coefficient on the same sets of data. Quaestiones Geographicae 30(2), Bogucki Wydawnictwo Naukowe, Poznań. 2011;87–93. 3 figs, 1 table.
    DOI: 10.2478/v10117-011-0021-1
    ISBN: 978-83-62662-62-3.
    ISSN: 0137-477X.

15. West DM. Digital divide: Improving Internet access in the developing world through affordable services and diverse content. 2015;3.
(Accessed on 13 June, 2018)
Available:https://www.brookings.edu/wp-content/uploads/2016/06/West_Internet-Access.pdf
16. Marta Guerriero. The impact of Internet connectivity on economic development in Sub-Saharan Africa. page 20. University of Birmingham; 2015.
(Access on 13 June, 2018)
Available:https://assets.publishing.service.gov.uk/media/57a0899b40f0b652dd0002f4/The-impact-of-internet-connectivity-on-economic-development-in-Sub-Saharan-Africa.pdf

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sciencedomain.org/review-history/28136