IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 09 Issue 03, March 2022
ISSN (Online) 2348 – 7968 | Impact Factor – 6.72
www.ijiset.com

Predicting Cost of ICT in Public Universities Using Regression Algorithm

David Odera[1], George Ogongo, PH.D [2], Dancan Okello[3]

[1] Computer Science, Tom Mboya University College, P. O. Box 199-40300, Homa-Bay, Kenya

[2] Mathematics, Tom Mboya University College, P. O. Box 199-40300, Homa-Bay, Kenya

[3] Computer Science, Tom Mboya University College, P. O. Box 199-40300, Homa-Bay, Kenya

ABSTRACT

Why do we continuously experience ICT budget shortfalls in our public universities in Kenya? There is growing need for institutions to be more knowledgeable about ICT expenses and budgets in public universities for them to understand ICT priorities in various institutions. This is very important because ICT supports services which improves overall organizational value proportions. ICT budgets in most universities have been growing exponentially and this leads to far reaching effect on the entire university budget. These budgets alongside other criteria largely informs the university funding board to apportion funds to universities. Therefore, a need to capture historical and existing ICT budgets in public universities in order to perform data mining and analysis is paramount. This study propose a data mining technique using Multiple Regression Algorithm in predicting ICT expenditure in Public Universities based on ICT costs and students population in Kenya. This technique would mine data which originate as a sample from public universities and collected in Ms. Excel sheet. The implementation logic would depend on developer's programming language of choice; however, Python Programming Language is preferable. The parameters like internet bandwidth, software licensees, hardware and cost of support, maintenance would be selected to conduct this research. Resultant prediction model can be used to identify cost of ICT expenditure and will also show relationship between location of university and cost of ICT.

Keywords: *Python-High level Programming Language, MOE-Ministry of Education, Naïve Bayes Classifier- Algorthm that assumes that presences of a particular feature in a class is unrelated to presences of any other feature, Regression Algorithm: models relationship between dependent and independent variables*

Purpose

This study seek to achieve its general objective of developing a predictive model for cost of ICT expenditure in a given period by focusing on the following specific objectives. To design predicting technique by applying regression mathematical model using Python language. To predict future cost of ICT expenditure using the above model developed in Python language

1. INTRODUCTION

A growing need to use data mining techniques in institution's data in order to understand various patterns and make adequate analysis for improved decision have gained popularity. This concept paper models a technique, which implements multiple regression in predicting cost of ICT expenditure in higher learning institutions. Osborne (2000) explains multiple regression as an algorithm that involves prediction and explanation. Multiple regression is a type of supervised machine learning where linear relationship between input and output variables are involved. This concept builds upon the various insights that were undertaken by various researchers in the business and information technology field to bridge the gap of strategic implementation of predictive systems in business to gain competitive advantage or to standardize operations. Olaniyi, Kayode & Jimoh (2011) attest to the fact that most researches have used regression to indicate if independent variable have a significant relationship with dependent variable, they have also used it to show strengths of independent variable's effect on dependent variable and finally in predictions.

1.1 Need of Study

Ministry of Education's Sessional Paper NO. 1 of 2019 on Policy Framework for Reforming Education and Training for Sustainable Development in Kenya highlights major challenges facing ICT integration in education and training such as inadequate ICT equipment; lack of or poor internet connectivity; unreliable power supply; inadequate ICT integration capacity among others. According to Ministry of Education (2019), those challenges can be addressed through certain policies regarding funding and provision of ICT in education in all respects as a national development priority. MOE also directs ICTs for education to allocate *specific* and adequate annual budget. Currently the government provides budgetary support to public universities in direct proportion to the total number of Full-Time Student Equivalent (FTSE) in each institution. However, specific reports containing allocations with respect to expenses such as ICT costs are lacking in the government data hub.

As noted by Kavulya (2006), two entities that need to be budgeted for are capital expenditures and operating budget. Capital expenditures include allocations for fixed assets such as new buildings, renovations, and ICT installation of automated systems. It also includes budgetary provisions for maintenance, replacement, repair, and renovation – and for investment in new and improved means of information access and delivery. According to The Kenya National ICT Master Plan 2014-2017, the government shall fund the foundational pillars through a re-focused expenditure-planning model. This technique can therefore be setup as part of data hubs that drive the e-Government systems and applications and the recurrent expenditure for operations. The allocating ministry and institution can then use it as a baseline when allocating ICT resources particularly in educational institutions such as universities.

The need of this study is multifaceted, universities may use the estimated costs associated with ICT in developing realistic budgets for ICT. Moreover, institutions would be more knowledgeable in understanding how others prioritize their ICT investments.

Dong et al. (2020), proposed a potential trend for online shopping data based on the linear regression and sentiment analysis. The naive Bayes (NB) classifier is applied to extract the sentiment orientation (positive or negative) from the Amazon product reviews. The sentiment orientation is quantified into 11 levels. Based on the linear regression and multiple linear regression models, they analyzed the three product datasets (microwave oven, baby pacifier, and hair dryer) to provide meaningful quantitative relationships between star ratings, reviews, and helpfulness ratings that will help the e-commerce company succeed in their online marketplace product offerings. Descriptive statistics was used to analyze the relationship between the three datasets and their variation trend. Olaniyi, Kayode & Jimoh (2011), studied regression analysis for use in stock price prediction. They used a data mining tool to uncover patterns and relationships and also extract value from database to predict. However, variables involved were all quantitative in nature.

Taboea, Salakoa, Tisonb, Ngonghala, Kakaï (2020) predicted COVID-19 spread in the face of control measures in West Africa. They formulated and used a deterministic compartmental model to predict the future course of the pandemic with and without currently implemented and additional control measures in West Africa. The deterministic -type model framework, where the total population (N) is subdivided in four categories: Susceptible (S), Exposed (E), Infectious asymptomatic (I_a), infectious symptomatic (I_s), infectious at treatment or isolation centers (I_c), and Recovered (R). That is, $N=S+E+ I_{a+} I_{s+} I_{c+} R$. This approach models the force of infection in a functional form $(1-\Psi)(\beta_a I_a + \beta_s I_s) /(S+E+I_a+I_s+R)$, where $\beta_k(k \in \{a,s\})$is disease transmission rate by individuals in the Ik class and $0\leq \Psi \leq 1$ is the percentage reduction in disease transmission due to public health control. This method and context in which its being applied is different from multiple regression formula. Šebalj, Franjković & Hodak (2017) proposed a shopping intention prediction using decision trees. They created a model that is able to predict shopping intention and classify respondents into one of the two categories, depending on whether they intend to shop or not. They used decision trees method in order to create a model with its several classification algorithms.

Krishnan & Patel (2020) developed a Regression Analysis of the probability of a recession and student loan debt utilizing data between 1993-2019. They trained a Long Short Term Memory (LSTM) model, an artificial recurrent neural network (a series of algorithms that connects underlying relationships in the data and can predict future

outputs), with the already available data points provided and predicted the future total GDP Values of 2019 through 2021.

Therefore, it's evident that little research have been done particularly in developing predictive models using regression and applying them in the analysis of both quantitative and qualitative variables.

1.1.1 The Gap

According to Bulcahnd, Kereteletswe, Castro & Molebatsi (2011) there is need to be in a position to justify to government about reasons why ICT expenditure is a very important part of ICT budget and overall university growth. They allude that anytime there is economic meltdown, ICT budget cuts are experienced. This means that most decision makers do not appreciate value-addedness from investing in ICT. This gap is attributed to lack of knowledge about ICT expenditures from various government institutions universities being some of them. Most of the time the cost of ICT expenditure is related to total amount of university's ICT budget. There is need to model a framework which can be used to obtain closer approximation in predicting expected ICT expenditure in given period for universities. A question on whether or not geographic location has a stronger or weaker relationship to cost of ICT expenditure is of significance, since geographic location is qualitative in nature and cost of ICT expenditure is quantitative. A search in literature database indicate a gap in predicting ICT costs, therefore need to conduct this study.

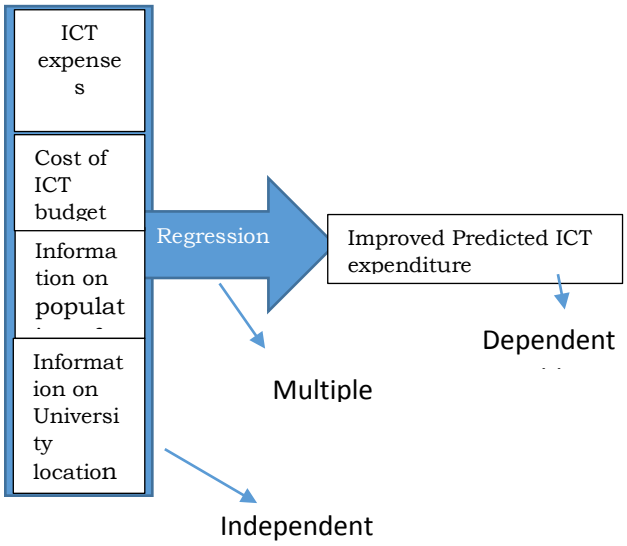

Fig 1: Block diagram showing relationship among variables

Based on the block diagram above, more than one independent variables would be used for analysis hence need to apply multiple regression algorithm. Multiple regression requires two or more predictor variables, and this is why it is called multiple regression. A dependent variable is modeled as a function of several independent variables with corresponding coefficients, along with the constant term.

Using Hardy (1993) descriptions, the estimated multiple regression equation for this study would take the following form

$$Ey = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \beta_0.$$

Here, β_i's (i=1,2…n) are the regression coefficients, which represent the value at which the criterion variable changes when the predictor variable changes. Y is the cost of ICT expenses (dependent variable), X_1 is cost of

hardware, X_2 is ICT budget, X_3 is student population and so on. β_0 is taken as intercept. Let's say the cost of ICT expenses would depend on various factors like cumulative cost of hardware, cost of software licenses, ICT budget, support/maintenance cost, student population and number of staff. Using these tests you can appropriate relationship among these factors.

Coefficient of determination(R Squared) formula would be used to analyze how differences in student population or ICT budget is explained in difference in cost of ICT expenditure. R-squared gives you the percentage variation in y explained by x-variables. R^2 is similar to correlation coefficient which gives the strength of regression between two variables. The formula is given as

$$R = 1 - \frac{sum\ squared\ regression\ (SSR)}{total\ sum\ of\ squares\ (SST)}$$

$$R = 1 - \frac{\sum\ (y_i - \hat{y}_i)^2}{\sum\ (y_i - \bar{y}_i)^2}$$

The sum squared regression is the sum of the *residuals* squared, and the total sum of squares is the sum of the distance the data is away from the mean all squared. As it is a percentage it will take values between 0 and 1.

Residual r =actual *y* value−predicted *y* value

1.1.2 Assumption

Not all independent variables will be used only relevant variables will be included in the model for it to be reliable. The model would be linear in nature. It's advisable to assume normality in multiple regression, hence variables would have a normal distribution.

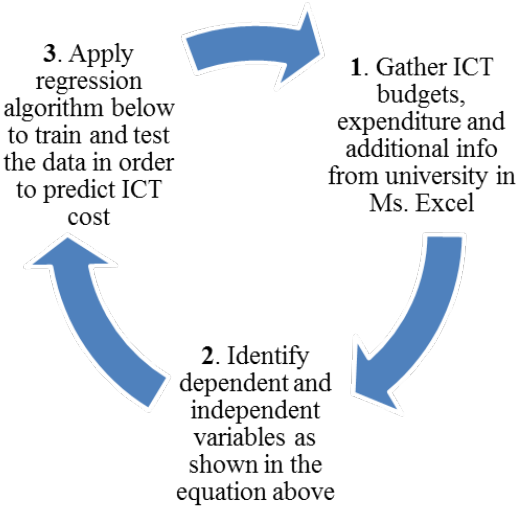

Figure 2: Flow chart of Predictive Model

2. METHODOLOGY

The study intends to use cross sectional survey design in determining the cost of ICT expenses in public universities. The study will sample 4 public Universities in Kenya out of a population of 40 which are registered. The study will

collect both primary and secondary data, and the quantitative data generated will be analyzed using descriptive statistics, which will include percentage distribution, mean and the frequency counts. The figure below illustrates the flow chart of a predictive model.

The model will be developed in Python language using regression mathematical formula. The relationship between the independent and dependent variables will explained through multiple regression. The process of developing the model would follow the order indicated below

Step 1: Import Libraries

```
import pandas as pd
import numpy as np
```

Step 2: Import dataset

```
data_df=pd.read_csv("/path of dataset")
data_df.head( )
```

Step 3: Define x and y i.e dependent and independent variables

```
x=data_df.drop(['dependent variable'] , axis=1).values
y=data_df['dependent variable'].values
```

Step 4: Split the dataset in training set and test set (we use sklearn.model_selection function for train and test set)

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3, random_state=0)
```

Step 5: Train the Model on the training set

```
from sklearn.linear_model import LinearRegression
ml=LinearRegression( )
ml.fit(x_train, y_train)
```

Step 6: Predict the test set results

```
y_pred=ml.predict(x_test)
ml.predict{([copy paste values of first row of y]
```

Step 7: Evaluate the model

```
from sklearn.metrics import r2_score
r2_score(y_test, y_pred)
```

Step 8: Plot the result

```
import matplot.pyplot as plt
plt.figure(fisize=(15,10)
plt.scatter(y_test, y_pred)
plt.xlabel('Actual')
plt.ylabel('Predicted')
plt.title('Actual vs Predicted')
```

Step 9: Predicted values

```
pred_y_df=pd.DataFrame({'Actual Value':'Predicted value':y_pred, 'Difference': y_test-y_pred })
pred_y_df[0:20]
```

3. CONCLUSION

The paper describes an algorithm based on regression mathematical notation that predicts the outcome of a variable representing cost of ICT expenses based on the values of other variables as indicated above. The technique enables analysts to determine the variation of the model and the relative contribution of each independent variable in the total variance. The regression analysis performed by the algorithm may help to show how much predictive information is associated uniquely with each predictor when you control for or partial out or "remove" any overlap or correlation with all other predictors.

4. REFERENCES

Aguinis, H. (2004). *Regression analysis for categorical moderators*. New York: Guilford Press.

Anderson, E. B. (2004). Latent regression analysis based on the rating scale model. *Psychological Science, 46*(2), 209-226.

Berk, R. A. (2003). *Regression analysis: A constructive critique*. Thousand Oaks, CA: Sage Publications.

Bulchand J., Kereteletswe O., Castro A., Molebatsi M. & Maedza M. (2011), "Justification Of ICT Expenditure: Input And Outputs" Conference Discussion Group Paper, University Of Las Palmas De Gran Canaria, Spain, Office Of The President (OP), Botswana Centre Algoritmi, Portugal, Ministry Of Transport And Communications – DIT, Botswana Ministry Of Education And Skills Development – IT Unit, Botswana, Ministry Of Education And Skills Development – Central Region, Botswana

Dong J., Chen Y., **Gu A.**, Chen J., Li L., Chen Q., Shujun L., & Xun Q. (2020), "Potential Trend for Online Shopping Data Based on the Linear Regression and Sentiment Analysis" Creative Commons Attribution License, https://www.hindawi.com/journals/mpe/2020/4591260/

Ghosal S., Sinha B., Majumder M., & Misra A. (2020), "Estimation of effects of nationwide lockdown for containing coronavirus infection on worsening of glycosylated haemoglobin and increase in diabetes-related complications: A simulation model using multivariate regression analysis" , https://doi.org/10.1016/j.dsx.2020.03.014

Hardy, M. A. (1993). *Regression with dummy variables*. Newbury Park, CA: Sage Publications.

Kavulya, J.M. (2006), "Trends in funding of university libraries in Kenya: A survey", Emerald Group Publishing, available at https://www.researchgate.net/publication/235253116_Trends_in_funding_of_university_libraries_in_Kenya_A_survey

Ministry of Information and Communication (MoIC) (2017), The Kenya National ICT Master Plan 2014-2017, Nairobi.

MINISTRY OF EDUCATION (2019), SESSIONAL PAPER NO. 1 OF 2019 on A Policy Framework for Reforming Education and Training for Sustainable Development in Kenya

Ministry of Information and Communication (MoIC) (2011), Draft Information and Communication Technologies (ICT) Sector Policy, Nairobi.

Olaniyi S. A, Kayode A. S. & R. G. Jimoh (2011) "Stock Trend Prediction Using Regression Analysis – A Data Mining Approach " ©2010-11 AJSS Journal.

Osborne, J. W. (2000). Prediction in multiple regression. Practical Assessment, Research, and Evaluation, 7(1), 2.

Published 2017 Engineering, Technology, Management and Tourism, DOI: https://doi.org/10.29352/mill0204.01.00155

Šebalj D., Franjković J. & Hodak K. (2017) "Shopping intention prediction using decision trees"

Taboe B. H., Salako V. K., Ngonghala C. N & Kakai R. G.(2020) "Predicting COVID-19 spread and public health needs to contain the pandemic in West-Africa" doi: https://doi.org/10.1101/2020.05.23.20111294

Yash P. & Pranav K, (2020). "A Regression Analysis of the probability of a recession and student loan debt utilizing data between 1993-2019," SocArXiv exnjd, Center for Open Science.