International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017), e-ISSN: 2455-2585
Volume 5, Issue 03, March-2019

FORECASTING UNION BUDGET OF INDIA BY LINEAR REGRESSION WITH STANDARD ERROR ADJUSTMENT

Mohd Ashar S.A. Khan[1], Prof. C.M. Mankar[2]

[1]*CSE Dept., SSGMCE, Shegaon, India,*
[2]*CSE Dept., SSGMCE, Shegaon, India,*

Abstract— An accurate budget forecast of any country before starting of any financial year is very important; it is a key factor in determining the budgetary allocation to various subjects. India is a vast country with nearly 1.2 billion people. Hence, it is difficult to forecast Indian budget quickly and accurately at the planning stage, when data, documentation, etc. are incomplete. The solution of this problem is addressed by our model. Using this model is an easy, fast and accurate way without having to prepare the data and other documentation first, even when just enough data, for example previous year's budget is given. The suggested solution using the model can be used to predict how much the Indian budget will be. The suggested model has passed statistical test of linear regression analysis. Forecasting budget using this model is easier, fast and accurate, so decision for making the necessary budgetary allocation can be done very quickly.

Keywords— Machine learning, Time series, Artificial intelligence, Linear regression, Predictive analytics.

I. INTRODUCTION

The Union budget of India is annual financial statement passed by the Parliament, approved by the President of India and presented by the Finance Minister. The budget is Annual Financial Statement of India. It forecasts the government revenues for the next financial year.

A. Need of Indian Budget

India is a very vast country with population more than 1.2 billion people. Union budget of India is approx. in thousands of billion Rs. Hence, it becomes very important owing to all these reasons that budget forecasting should be done before-hand so as to get a rough estimate of revenues which will be generated and thus utilized appropriately.

In this paper we have tried to forecast the revenue element of the union budget of India with the help of predictive analytics and data science. The machine learning algorithm which is used in this paper is Linear Regression model.

B. Concept

The goal and usage of Machine Learning is to build new and leverage existing algorithms to learn from datasets, in order to build generalizable models that give accurate predictions, or to find patterns with similar data. ML algorithms are heavily based on statistics and mathematical optimization [1].

The advantage of budget data facilitates analysis to categorize the optimum combination of outcome of risks and assets overtime. Under these conditions, the subjective judgment of decision makers is a crucial factor in making accurate forecasts to manage financial needs [6].

Methods of making forecasts fall into two categories: causal methods and time series methods [8]. In this paper, we are interested in making forecasts using the time series methods.

II. LITERATURE SURVEY

A review of previous research related to present topic is as follows:

A. Authors

1) Alvin C. Rencher, G. Bruce Schaalje [9].
2) Ouahilal Meryem, Jellouli Ismail, El Mohajir Mohammed. 2014 [8].
3) Cheng Lin, Fan Yan. 2015 [7].
4) Azamat Kibekbaev, Ekrem Duman. 2015 [6].
5) Sphurti S. Arage, Nagaraj V. Dharwadkar. 2017 [5].
6) Paikun, Trihono Kadri, Ria Dewi Hudayani Sugara. 2017 [4].
7) Howard J. Seltman. 2018 [3].
8) Lijuan Wang, Guodong Li. 2018 [2].
9) Kapil Bakshi, Kiran Bakshi. 2018 [1].

In the literature there are only a few empirical studies on financial prediction.

Regression analysis is a statistical technique used to find or estimate relationships among variables. It is used when you want to predict a relationship between a dependent variable and one or more independent variables [4].

Regression analysis are of two types viz. nonlinear regression and linear regression. In this work we will use linear regression technique. This is why, we wanted to provide some articles from the literature related to financial estimation/forecasting by regression models.

Bakshi and Bakshi [1] in their paper has tried to discuss artificial intelligence and machine learning, their co relation. Various approaches of machine learning with their practical use cases.

Wang and Li [2] dealt with economic forecast and feasibility to build a financial centre in Urumqi using linear regression model of predictive analytics.

Seltman [3] described in his book about linearity of data, standard deviations and its adjustments statistically and graphically.

Paikun, Kadri and Sugara [4] showed a computational model of forecasting budget of housing and building project using linear regression techniques.

Arage and Dharwadkar [5] in their paper has estimated cost of construction project using linear regression model.

Kibekbaev and Duman [6] proved in their paper that for income prediction topics the best algorithm in machine learning paradigm would be linear regression technique.

Lin and Yan [7] discussed data mining techniques and superiority of linear regression in that.

Ouahilal, Jellouli and El Mohajir [8] showed us that for time series forecasting the best model to use is linear regression computational model.

Rencher and Schaalje [9] in their book discussed linear regression with each and every formula and proof in vast details. The mathematical and statistical aspect of this paper derives greatly from this work.

III. PROBLEM STATEMENT

The formulation of the problem is the basis of problem solving in the form of questions or desires of a thing based on the assumptions of questions that must be answered. For example, how to calculate an easy and practical residential cost budget plan without having any special skills and sufficient experience in the art. The formulation of the problem is a straight line of red thread in the research so that the research focus on the formulation of the specified problem [4].

India is a very large democracy and hence Indian budget hugely affect the people. Why budget forecasting is important? -

- Economic growth
- Businesses get direction
- Reduces disparities
- The budget caters to certain government objectives
- Taking care of Public Sector Undertakings

Also, for long term planning it is very crucial to have future figures.

Owing to all the above mentioned points, it becomes extremely important to come beforehand the forecasts and forecasts of budget to have a better idea at policy formation. Hence in this paper, we will try to explore the ways and develop a working computational model to forecast Union of India Budget.

IV. DATASET PREPARATION

A. *Data Collection*

Based on literature review and hypotheses planned to analyse the identified problem, data required, that is data of Union Budget of India, data obtained from the website of Indian Budget, www.indiabudget.gov.in [10].

The data is contained in various files on the website. Hence appropriate data is gathered and collected in table format. After going through data pre-processing, selection of variable was done in order to remove redundant or irrelevant characteristics. The main aim of attribute selection is to improve the performance of regression model.

B. *Analysis of Data*

In this section we have already discussed data collection, then predict the variables x (independent variable) which may affect the variable y (dependent variable). Based on previous research findings and hypotheses that the variables that will affect y (union budget of India) is x (financial year). So that the data needed for analysis is revenue Budget of Union of India which is taken from the official website of Indian Budget [10].

V. METHODOLOGY

A. *Time Series Forecasting*

There are several methods of making forecasts, but they all fall into two categories: causal methods and time series methods [8].

Time series is based on chronological sequence of observations carried out on a particular variable. Generaly the observations are taken at regular time intervals (days, months, years), but data sampling could be irregular. Analysis of time series consists of two steps: (1) developing a model representing time series. (2) using this very model to predict and forecast future values [8].

In this paper, we are interested in making forecasts using the time series methods.

B. *Predictive Data Mining*

There are two types of machine learning algorithms; supervised machine learning algorithms and unsupervised machine learning algorithms. Machine learning algorithms that learn from input/output pairs are called supervised learning algorithms because a "supervisor" provides guidance to the algorithms in the form of the desired outputs for each dataset they learn from. All the supervised machine learning algorithm can be seen as either classification or regression [1].

In regression, the objective is to predict a continuous number with a value. Predicting a person's annual income from their education, their age and where they live is an example of a regression task [1]. In social sciences, business and engineering, linear models are quite useful in planning stages of research and analysis of the data [9].

Although the linear regression is one of the most simple regression form, but linear regression is more often used than other complex regression method and gives better approximation, so it is widely used in actual calculation [7]. Simple linear regression has very good performance, which indicates that income can also be calculated by the linear regression [6].

In linear regression relationship between dependent variable and independent variable is modelled using linear predictor functions. Mathematically, the expression can be shown as y = a + b* x where, y = Forecasted Budget (i.e. Dependent Variable), a = Intercept, b=Slope, x= Year (i.e. Independent Variable) [1][2][5][7][8][9].

C. *Standard Error*

The error model that we developed is for each x, the values of y corresponding to that of x, their distribution is Gaussian with a spread, σ^2. We can make an forecast of σ^2 from the data. The error model includes the assumptions of "Normality" and "equal variance" and assumption of "fixed-x". The assumption of "fixed-x" states that the explanatory variable(year in this case) is measured without any error. In our dataset as the time period is continuous and the model is time series, we can conclude that "x" is fixed and error free. [3]

The four little Normal curves represented in the Fig. 1 are the Normally distributed outcomes (Y values) at each of four fixed x values. The four Normal curves with same spreads represents the equal variance assumption. And the four means of the Normal curves along a straight line represents the linearity assumption [3].

Standard errors, σ^2 are estimated standard deviations of the corresponding sampling distributions. A positive residual indicates a data point higher than expected, and a negative residual indicates a point lower than expected [3].

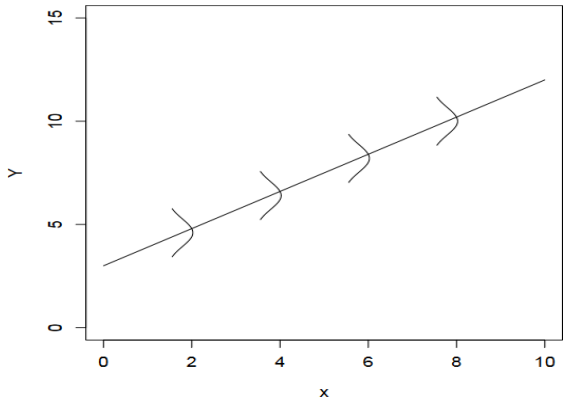

Fig. 1. Normally distributed values of Y

For adjusting standard error in our model, we have calculated it and presented this standard error along with the forecast as a range where the forecasted budget would fall. The standard error is first subtracted from the forecasted value to give the lower bound of the budget and then the standard error is added in the forecasted value to give the upper bound of the range of the forecasted budget for a given future year.

VI. RESULT AND DISCUSSION

A model is a mathematical construct that represent the mechanism that generated the observations at hand. The postulated model may be an idealized oversimplification of the complex real-world situation, but in many such cases, empirical models provide useful approximations of the relationships among variables. These relationships may be either associative or causative [9].

The solution to solve the problem is to create a computational model, an application software. This system is built to forecast the union budget of India. As a result of the implementation and evaluation that has been done, the system can help to forecast revenue budget for future.

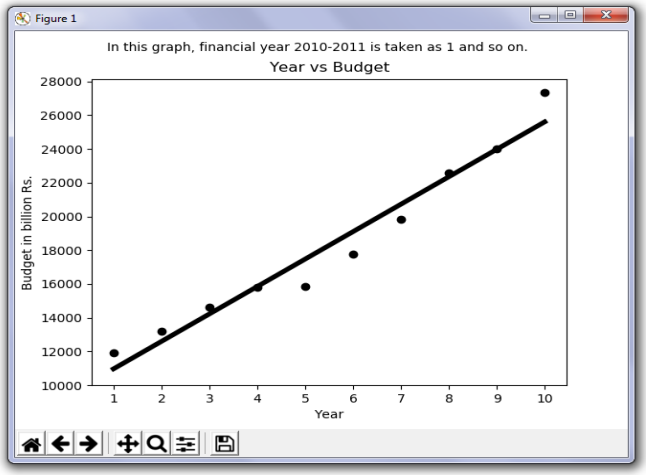

Fig. 2. Linear regression graph

For efficiency and effectiveness last 10 years of data is taken, after preparing various graphs to check linearity in data. The coefficient of determination (R^2) of the developed model is 0.9578940082369579. This shows that the relationship between the budget i.e. dependent and year i.e. independent variables of the developed model are good. For better fitting of model coefficient of correlation (R) should lie between -1 to 1 & coefficient of determination (R^2) should lie between 0 and 1 [5].
R-square which is also known as square of the multiple correlation coefficients and the coefficient of multiple determinations, give the square of the correlation between the response values and the predicted response values [6].

A. *Regression Analysis*

1) Descriptive statistic: Statistic description is done to test the normality of data before doing further analysis, normality data test can be done with normal plot regression, there are diagonal line and dots.

In the fig. 2, we can see that the assumptions of linearity seems plausible because we can draw a straight line from bottom left to top right going through the center of the points [3].

From the test results of the graph method by looking at the spreading point on the diagonal axis of the graph. Basic decision-making; 1) if the data spreads around the diagonal line and follows the direction of the diagonal line then the regression model meets the assumption of normality; 2) if the data spreads far from the diagonal line and does not follow the diagonal line direction then the regression model does not meet the assumption of normality [4]. Linear regression is the best prediction algorithm to provide the time series forecasting. [8]

B. *Model Description*

In TABLE I, various metrics that has been generated by this linear regression model is given:

TABLE I
VALUES GENERATED BY LINEAR REGRESSION MODEL

Metrics	**Values**
Intercept	9349.123333333331
Slope	1625.24230303
Mean Absolute Error	778.0847999999999
Mean Square Error	957892.182460241
Root Mean Square Error	978.7196648991176
Co-efficient of Determination, R^2	0.9578940082369579
R	0.9787205976359943
% Accuracy	95.78940082369579 %
Standard Deviation, Sy.x, σ^2	1094.2418508151209

C. *Validation Test*

Model that has been developed can be used to do testing of factual data to know whether the model can be used and how accurate it is, validation test is shown in TABLE II.

A residual (difference between actual and calculated budget) is the deviation of an outcome from the predicated mean value for all subjects with the same value for the explanatory variable.

D. *Forecasted values*

TABLE III shows the forecasted union budget of India by linear regression model of predictive analytics for the given future financial years and also range within which the budget may fall (after adjusting standard error).

Model forecasting of the budget is: Y = 1625.24230303X + 9349.1233 . The output gives accuracy of the forecasting model of union budget of India as 95.78940082 % with error distribution of 1094.24185082 and average percentage difference between actual and calculated budgets is -0.0069244765. Model we propose will give accurate forecasts of the union of India budget.

TABLE II
PERCENTAGE DIFFERENCE IN ACTUAL BUDGET AND CALCULATED BUDGET

Financial Year	**Total Budget in Billion Rs. (A)**	**Calculated Budget in Billion Rs. (B)**	**% (A-B)**
2010-2011	11908.99	10974.365636363635	7.85
2011-2012	13203.55	12599.607939393938	4.57
2012-2013	14613.83	14224.85024242424	2.66
2013-2014	15786.18	15850.092545454543	-0.4
2014-2015	15858.29	17475.334848484847	-10.2
2015-2016	17776.04	19100.57715151515	-7.45
2016-2017	19840.89	20725.819454545453	-4.46
2017-2018	22571.29	22351.061757575757	0.98
2018-2019	23991.47	23976.30406060606	0.06
2019-2020	27329.03	25601.546363636364	6.32

TABLE III
FORECASTED VALUES AND RANGE OF BUDGET OF UNION OF INDIA AGAINST FINANCIAL YEARS

Financial Year	**Forecasted Budget in Billion Rs.**	**Forecasted Budget Range in Billion Rs.**
2020-2021	27226.78866667	[26132.54681585] TO [28321.03051748]
2021-2022	28852.0309697	[27757.78911888] TO [29946.27282051]
2022-2023	30477.27327273	[29383.03142191] TO [31571.51512354]
2023-2024	32102.51557576	[31008.27372494] TO [33196.75742657]
2024-2025	33727.75787879	[32633.51602797] TO [34821.9997296]

So result shows that proposed model works efficiently and it can be used to forecast budget of India.

VII. CONCLUSION

The input dataset contains 10 years of union budget of India. When we plotted graph for this dataset, it shows linear relationship between year and budget. As the input dataset shows linear nature, we proposed linear regression model.
The model is as follows: Y = 9349.12 + 1625.24X
The accuracy of budget forecasting using this model is 95.79%, with an average error of -0.007%.
As R lies between -1 to 1 and R2 lies between 0 to 1, it shows that proposed model better fits the data. The accuracy of proposed model is greater than 97%. Based on statistical test of linear regression analysis yielded a model, this model is a solution to forecast the union budget of India, because it is simple, very easy, fast and accurate, this model is easy to use. Thus prediction of future budget of union of India can be done using dataset of last 10 years of union budget. So with our proposed model we have forecasted the future budget of India.

REFERENCES

[1] Kapil Bakshi, Kiran Bakshi. *Considerations for artificial intelligence and machine learning: Approaches and use cases.* IEEE Aerospace Conference. 2018

[2] Lijuan Wang, Guodong Li. *Economic Forecast for Urumqi to Build a Regional Financial Center.* 10th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA). 2018

[3] Howard J. Seltman. *Experimental Design And Analysis.* http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf . 2018

[4] Paikun, Trihono Kadri, Ria Dewi Hudayani Sugara. *Estimated Budget Construction Housing Using Linear Regression Model Easy And Fast Solutions Accurate.* International Conference on Computing, Engineering, and Design (ICCED). 2017

[5] Sphurti S. Arage, Nagaraj V. Dharwadkar. *Cost estimation of civil construction projects using machine learning paradigm.* International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC). 2017

[6] Azamat Kibekbaev, Ekrem Duman. *Benchmarking Regression Algorithms for Income Prediction Modeling.* International Conference on Computational Science and Computational Intelligence (CSCI). 2015

[7] Cheng Lin, Fan Yan. *The Study on Classification and Prediction for Data Mining.* Seventh International Conference on Measuring Technology and Mechatronics Automation. 2015

[8] Ouahilal Meryem, Jellouli Ismail, El Mohajir Mohammed. *A comparative study of predictive algorithms for time series forecasting.* Third IEEE International Colloquium in Information Science and Technology (CIST). 2014

[9] Alvin C. Rencher, G. Bruce Schaalje. *Linear models in statistics.* Second edition. Published by John Wiley & Sons, Inc.

[10] Datasets. https://www.indiabudget.gov.in