Proceedings of 14th IOE Graduate Conference
Peer Reviewed
Year: 2023 Month: December Volume: 14
ISSN: 2350-8914 (Online), 2350-8906 (Print)

Machine Learning Techniques for Cost Estimation of Road Construction in Nepal

Raj Basnet [a], Samrakshya Karki [b]

[a, b] Department of Civil Engineering, Pulchowk Campus, IOE, Tribhuvan University, Nepal
[a] basnet.raaz21@gmail.com [b] karkisamrakshya@gmail.com

Abstract

Cost estimation is one of the important parameters for successful completion of construction project during every stages. For developing countries like Nepal, cost estimation is one of the challenging factors due to the lack of proper database management system and unavailability of historical data. Despite such, primary data were collected by surveying, which were then applied to generate a cost estimation prediction model with the help of supervised ML techniques. Algorithms such as ANN and XGBoost were applied successfully. The percentage accuracy obtained was 80.86% using ANN and 71.26% using XGBoost. Hence, the model generated using ANN was accepted, therefore ANN is a better ML algorithm for the cost prediction of road projects for developing country like Nepal.

Keywords

Cost Estimation, Road Projects, Artificial Neural Network, Extreme Gradient Boosting

1. Introduction

The construction industry is an important parameter for the economic enhancement of overall country, thus making road construction a necessity to connect the nook and corners of the entire region. In developing countries like Nepal, road transportation is the primary mode of transportation. The development of highway in developing countries such as Nepal is a prime case scenario [1]. Construction of interconnected roads and highways is essential and at the same time, is costly and time-consuming. The first phase of the project, the conceptual phase, where the project are established with an identification of sponsor, is the first cost-related phases ([2, 3]). The little information depicts uncertainty, but minimization of such uncertainty for completion in estimated time, cost and specifications marks project success [4].

The major part of construction projects is cost estimate, which is one of the main criterion's for initial decision-making and is prime factor to be considered during construction process [3]. The accuracy in the cost estimation is which affects the overall profitability of the construction projects at the tender and also the completion phase [4]. The early stages mark the difficulty, but when estimated correctly leads to right decisions [1]. The main problems during cost estimation leading to cost under-run and overrun are the lack of road cost database, poor management of preliminary data, proper consideration in estimation techniques, and exclusion of uncertainty [3].

Thus, there is a very important need for appropriate cost estimation techniques which propose suitable model development, for which, artificial intelligence comes in handy. Various studies use the regression method, but cannot handle complex parametric relationships [4]. Machine learning and data mining have become a part of our daily lives [5]. With various research in artificial intelligence (AI), it has been established as the fact that its use will be a supreme tool for the development of cost-estimation models in various highway infrastructures [1]. The use of machine learning in making preliminary estimates would significantly reduce the time and, consequently, the cost of data processing.The linear relationship between the project cost and uncertainty in traditional approach of cost estimation is what would reduce the precision and accuracy of the results [6] thus, there is the introduction of ML algorithms.

2. Literature Review

2.1 Cost Estimation Methods

Among various other parameters for the project succession, cost is one of the important factors which is considered the main highlight during preliminary and conceptual stage of the project as the key to further planning and feasibility studies [3]. Among various studies, the traditional approach to cost estimation is regression analysis which falls sort when presented with large number of variables [7].

A cost estimation model for building projects using various AI techniques were developed, where in most of the cases Extreme Gradient Boosting (XGBoost) outperformed other machine learning (ML) algorithms [4, 8, 9, 10, 11].

The most common AI techniques were reviewed for cost modelling to develop a total of 20 AI techniques for parametric cost estimation and highlighting the future trends [12]. A perspective of the cost regarding highway into terms of owners and contractors were provided to select the key cost drivers and estimate cost of highway construction projects, out of which XGBoost came in upper hand [1] .

ANN was used for cost estimation in road construction emphasizing choosing of predictors as the most important part of the modelling with neural networks where, general regression neural network (GRNN) stood out in accuracy with shortest analysis run time [3]. The ROCKS database were used to collect data of 315 projects to estimate highway cost using multi layer perceptron (MLP) with back propagation [2]. The total of 75 highway projects in Egypt were studied for developing parametric model for conceptual cost estimate of highway projects using supervised neural network MLP with back propagation [13]. A three layered NN with back-propagation had been used to model the parametric cost estimation of highway projects in Canada where a major of three approaches namely, back propagation, simplex optimization and genetic algorithm (GA) were used [7]. The use of NN with back-propagation with nine units in input layer, the project's characteristic vector, were used for the cost estimation of highway engineering yielding an error of less than 5 percentage suggesting the better use of NN rather than traditional regression analysis due to their non-linear relation between project cost and uncertainty [6]. The quantity based automated estimating system was developed using quantity estimating models on the cost data [14]. The cost estimation model using case-based reasoning was developed, which was processed with the application of rough set theory, thus combining rough set theory, case-based reasoning and genetic algorithm for better prediction [14]. The hybrid CBR model with analytic hierarchy process was developed for increasing the accuracy and time reduction during the early stage [15]. The total of 131 sets of road projects were used in Palestine for early cost estimation of road construction using multiple regression techniques developing ten regression models [16].

An ensemble learning model was developed to predict the unit bids of resurfacing highway projects, in addition to baseline Monte Carlo simulation and a multiple regression model for comparison [17]. An ensemble model with the use of ANN, SVM, and RF were built using k-fold cross validation to optimize the bias-variance trade-off for the conceptual estimation of construction duration and cost of public highway [18]. The database of 234 projects were used in South Korea to predict project cost overrun levels in bidding stage using ensemble model with a 10-fold cross validation [19]. Techniques such as text mining, numerical data, and ensemble classifiers were used to predict construction cost overruns using the projects characteristics. In all above cases, an ensemble model performed better result than other ML algorithms [20].

2.2 Artificial Neural Network

Artificial neural network (ANN) is a type of machine learning algorithm, that imitates the human brain and its biological function acting as a stimulus, is one of the popular methods used in various fields. A neural network consists of three layers input, hidden, and output. There will be huge reduction in time and cost with the use of artificial neural networks while making preliminary estimates. In case of provision of large data set, the application of ANN has always been found of good solution. Among several models, Multi-layer perceptron (MLP) with a back-propagation algorithm is of utmost use due to its easiness and accuracy. There has been vast distributed structure while using NN, thus increasing the computing power and ability to learn [21].

For this research, a multilayer feedforward neural network (MLFFNN) is used [22]. The feedforward network consists of the input vector (x), a weight matrix (W), a bias vector (b), and an output vector (Y) that can be formulated as [22]:

$$y = f(Wx + b) \tag{1}$$

Where $f()$ refers to a nonlinear activation function

Due to the nature to solve complex problems, MLFFNNs are immensely popular, however the need for long training time also has been a problem [21].

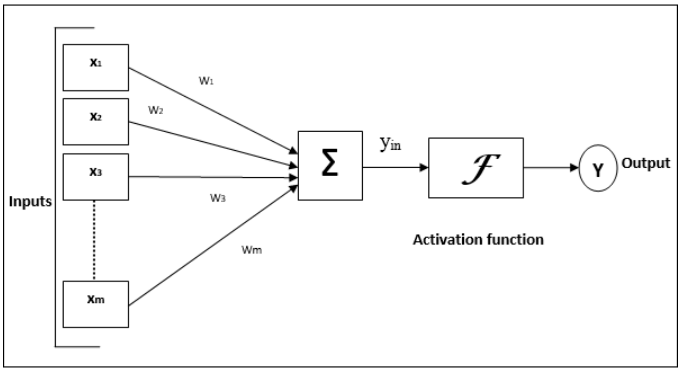

Figure 1: An Architecture of Neural Network

2.3 Extreme Gradient Boost (XGBoost)

Extreme gradient boost (XGBoost) is one the popular machine learning algorithms that is quick and is one of the tree boost methods. Recently, various engineering problems are being solved with use of the XGBoost algorithm [4]. The boosting algorithm's concept uses an iteration process for learning the functional relationship between the target and predictor values. Variants of the model have been applied to problems such as classification and ranking [5]. The use various modeling techniques such as multiple-regression analysis (MRA), artificial neural networks (ANN), and extreme gradient boosting (XGBoost) were done, where the final comparison yields XGBoost as the better cost estimator [1].

The general equation for XGBoost algorithm for cost estimation of road construction projects is as follows [4]:

$$Y = \phi(x) = \left[\sum_{i=1}^{n} L(y_i, p_i)\right] + \gamma T + 1/2\lambda 0^2_{value} \tag{2}$$

Where, T is the number of terminal nodes, γ is the user-definable penalty, and λ is the regularization parameter.

The loss function which is calculated as:

$$L(y_i, p_i) = \left[y_i \log p_i + (1 - y_i) \times \log(1 - p_i)\right] \tag{3}$$

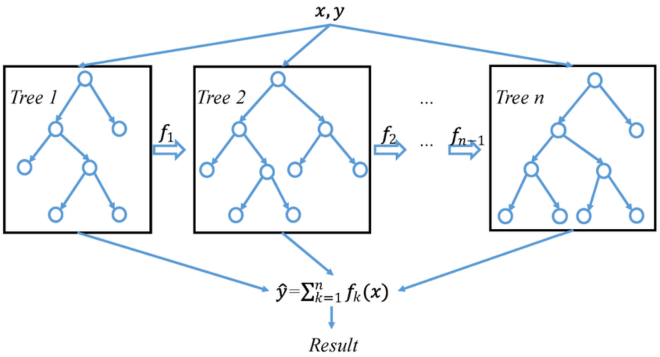

Figure 2: An Architecture of XGBoost Algorithm

3. Research Methodology

With thorough literature review, cost factors of highway projects were finally extracted from [22]. Data regarding the cost of road projects were collected from various sources such as renowned contractors, Department of Roads, consultancies and some articles. The validated factors were then selected via. feature engineering with the use of random forest (RF). Then the data were presented as per the required need of the model with change of various variables to binary features. Finally, the model was built as per the collected data and compared with the help of mean absolute percentage error (MAPE).

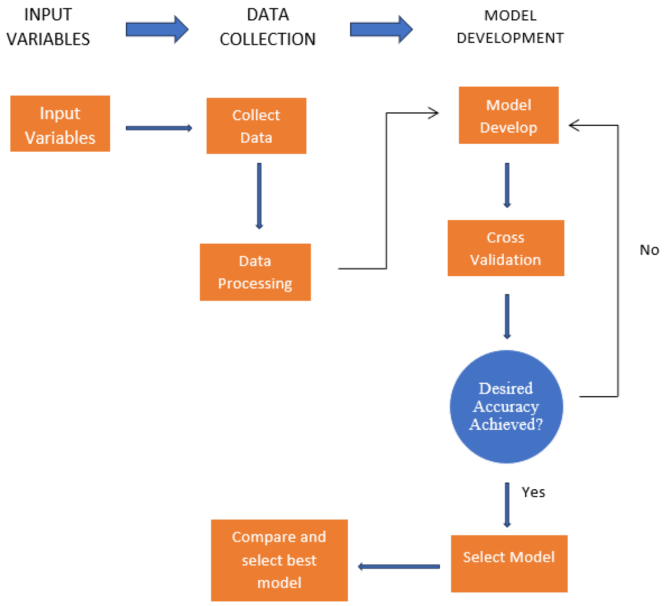

Figure 3: Flowchart of Research Methodology

3.1 Data collection and Feature Selection

The required data were collected from various public government sectors, renowned construction companies, consultants and some from previous articles. It was stated that the ratio of sample size to variables be 5 to 10 times as of that of variables [23, 24].

From the obtained parameters, feature selection was done using random forest (RF) which yielded in reduction of parameters giving more accuracy to the model.

The data was then pre-processed before feeding to the model by following steps [22]:

1. The location i.e., district was converted into accessibility factor

Table 1: Value for vehicle accessibility and environment factor

Vehicle Accessibility	Grade	Coefficient / Value
Very Difficulty	A	1.20
Difficult	B	1.10
Medium	C	1.05
Easy	D	1.00

2. As per the 2014/15 consumer price indes (CPI) provided by Nepal Rastra Bank (NRB), the actual cost were changed into base year cost which is given as:

$$\text{Base year cost} = \frac{\text{Actual Construction Cost}}{\text{Averaged CPI}} \times 100$$

3. Finally, various categorical variables were transformed into binary parameters (0,1).

3.2 Model Development

Model was developed using Python with its various other parameters and developed for ANN and XGBoost. ANN model was developed using Rectified Linear Unit (ReLU) function as activation function. The epoch list was considered as 50, 100, 500, and 1000. 3, 6, and 9 were considered as hidden list for trials with learning rates 0.1, 0.01, and 0.001. Similarly, XGBoost was developed with parameters like maximum depth as 10, 20, and 30, number of estimators as 50, 100, and 150 and learning rate as 0.1, 0.01, and 0.001. In both the model, the model with least mean absolute percentage error (MAPE) was selected with the best model architecture, which in turn gave the architecture with best accuracy and ultimately the model with best accuracy.

4. Result and Discussion

4.1 Variable Identification

After proper literature review and guidance from the previous researcher, input variables were finalized. The variables were categorized into three types which were all obtained from [22]:

Table 2: Categorical Variables

S.No.	Parameters
1	project Scope
2	Administrative Classification
3	Project Location
4	Terrain Type
5	Pavement Type
6	Flexible Pavement Type

Table 3: Binary Variables

S.No.	Parameters
1	Subgrade Works
2	Subbase Course Works
3	Base Course Works
4	Surface Course Works
5	Surface Drainage Works
6	Road Furniture Works
7	Utility Relocation Works
8	River Bank Protection Works
9	Check Dams
10	Retaining pr Brease Walls
11	Slope Protection Works
12	Bio Engineering Works
13	Footpath Works

Table 4: Categorical Variables

S.No.	Parameters
1	Length of Road
2	Width of Carriage Way
3	Percentage of Flexible Pavement
4	Sub base Course Thickness
5	Base Course Thickness
6	Surface Course Thickness
7	Year of Construction
8	Construction Duration

4.2 Data Collection and Feature Selection

The data from 107 road projects were collected from various sources which were classified as per some of the parameters. As per administrative classification, 34 were national highway, 39 were feeder road, 21 were district road and 13 were urban road. As per terrain, 38 were mountainous, 28 were rolling, 12 were steep and 29 were plain. As per pavement type, 5 were earthen, 57 were flexible, 32 were gravelling, 6 were partly flexible / rigid, and 7 were rigid. As per surface type, 38 were asphalt concrete, 19 were DBSD, 21 were otta seal, 6 were premix and 23 were none.

After data collection, feature selection was done using RF where the parameters later obtained were length, width, duration, scope, administrative classification, sub-base course thickness, base course thickness, surface course thickness, percentage of flexible pavement, location and check dam works.

4.3 Result of the model

After the feature selection, 14 input features were identified, which was used to model the two AI algorithms. ANN used rectified linear unit (ReLU) as activation function with epoch 1000, learning rate 0.001 giving MAPE of 19.14% during testing respectively with the best architecture of 14-3-1. Similarly, hyper-parameter tuning was done in XGBoost algorithm with learning rate 0.01, maximum depth 10 and number of estimators 150 yielding MAPE of 28.74%.

Both the above models generated an MAPE below 30% which as found acceptable as shown in the paper [3]. Also, result regarding the MAPE can be considered upto a margin of 50% [2, 4].

5. Limitation

This study was supposed to develop the model after all the hyper-parameter tuning, but due to time constraint, only some of the parameters like hidden layers, epoch and learning rate were tuned properly. The variation in hyper-parameter tuning is a prime factor to be considered during further research and enhancement along with different layers and their variations leading to possible up soar in reliability and accuracy of the model.

Also, only 107 data of road projects including large variations in the cost, were collected which is considered not good for AI modelling. Thus, researcher should consider to filter the data more precisely.

6. Conclusion

Construction industry, an important criterion in the overall economic development, plays a vital role in which cost estimation is also an important parameter. In developing countries like Nepal, cost estimation is very important to estimate the future expenses of the country to predict the budget accurately and is very important during bidding phases, to minimize the random bidding amount, ultimately emphasizing the success of the project. In this research, the factors affecting the cost estimation were collected from previous research paper and data were collected from DOR, renowned contractors, consultants and some from previous research. The total of 107 road projects data were collected. and transformed from categorical variables into binary variables for easy understanding by ML algorithms. A feature selection was done with the use of RF which limited the input parameters to 14 from overall 44 parameters. The data were then added to generate a cost estimation prediction model with the help of supervised ML techniques. Algorithms such as ANN and XGBoost were applied successfully which were compared based on MAPE. The percentage accuracy obtained was 80.86% using ANN and 71.26% using XGBoost. Hence, the model generated using ANN was accepted, therefore ANN is a better ML algorithm among the two for the cost prediction of road projects for developing country like Nepal. Despite the accuracy of both the models in the range, further enhancement can be done with increase of additional data along with changes required for the feature selection. The ML algorithms work more properly with more extreme hyper-parameter tuning which may increase the accuracy of the model, hence its reliability.

7. Acknowledgement

The authors express sincere gratitude towards Asst. Prof. Mahendra Raj Dhital for his immense help during this research work.

References

[1] N. Simić, N. Ivanišević, Đ Nedeljković, A Senić, Z Stojadinović, and M. Ivanović. Early highway construction cost estimation: Selection of key cost drivers. *Sustainability*, 15, 2023.

[2] J. Sodikov. Cost estimation of highway projects in developing countries: artificial neural network approach. *Journal of the Eastern Asia Society for Transportation Studies*, 6:1036–1047, 2005.

[3] K. Tijanić, D. Car-Pušić, and M. Šperac. Cost estimation in road construction using artificial neural network. *Neural Computing and Applications*, 32:9343–9355, 2020.

[4] Z. H. Ali, A. M. Burhan, M. Kassim, and Z Al-Khafaji. Developing an integrative data intelligence model for construction cost estimation. 1022.

[5] T. Chen and C Guestrin. Xgboost: Reliable large-scale tree boosting system. pages 13–17, 2015.

[6] DUAN Xiao-chen. Application of neural network in the cost estimation of highway engineering. *Journal of Computers*, 5, 2010.

[7] T. Hegazy and A. Ayed. Neural network model for parametric cost estimation of highway projects. *Journal of construction engineering and management*, 124(3):210–218, 1998.

[8] T. Q. D. Pham, T. Le-Hong, and X. V. Tran. Efficient estimation and optimization of building costs using machine learning. *International Journal of Construction Management*, 23(5):909–921, 2023.

[9] O Alshboul, A Shehadeh, G Almasabha, and A. S. Almuflih. Extreme gradient boosting-based machine learning approach for green building cost prediction. *Sustainability*, 14, 2022.

[10] Y. R. Wang, C. Y. Yu, and H. H. Chan. Predicting construction cost and schedule success using artificial neural networks ensemble and support vector machines classification models. *International Journal of Project Management*, 30:470–478, 2012.

[11] D. Chakraborty, H. Elhegazy, H. Elzarka, and L Gutierrez. A novel construction cost prediction model using hybrid natural and light gradient boosting. *Advanced Engineering Informatics*, 46, 2020.

[12] H. H Elmousalami. Artificial intelligence and parametric construction cost estimate modeling: State-of-the-art review. *Journal of construction engineering and management*, 146(1), 2020.

[13] K. Adel, A. Elyamany, A. M. Belal, and A. S Kotb. Developing parametric model for conceptual cost estimate of highway projects. *International Journal of Engineering Science*, 6:1728–1734, 2016.

[14] J. S. Chou, M. Peng, K. R. Persad, and J. T qnd O'Connor. Quantity-based approach to preliminary cost estimates for highway projects. *Transportation Research Record*, pages 22–30, 2006.

[15] S. Kim. Hybrid forecasting system based on case-based reasoning and analytic hierarchy process for cost estimation. *Journal of Civil Engineering and Management*, 19:86–96, 2013.

[16] I. Mahamid. Early cost estimating for road construction projects using multiple regression techniques. *Australasian Journal of Construction Economics and Building*, 11(4):187–101, 2011.

[17] Y. Cao, B. Ashuri, and M. Baek. Prediction of unit price bids of resurfacing highway projects through ensemble machine learning. *Journal of Computing in Civil Engineering*, 32, 2018.

[18] B. Mohamed and O. Moselhi. Conceptual estimation of construction duration and cost of public highway projects. . *J. Inf. Technol. Constr*, 27(29):595–618, 2022.

[19] H. Moon, T. P. Williams, H. S. Lee, and M. Park.). predicting project cost overrun levels in bidding stage using ensemble learning. *Journal of Asian Architecture and Building Engineering*, 19(6):586–599, 2020.

[20] T. P. Williams and J. Gong. Predicting construction cost overruns using text mining, numerical data and ensemble classifiers. *Automation in Construction*, 43:23–29, 2014.

[21] A. N Sharkawy. Principle of neural network and its main types. *Journal of Advances in Applied & Computational Mathematics*, 7:8–19, 2020.

[22] B. Acharya and S Karki. Artificial neural network for cost estimation of road projects in nepal. *IOE Conference*, 2022.

[23] J. C. Nunnally. An overview of psychological measurement. clinical diagnosis of mental disorders. pages 97–146, 1978.

[24] H. E Tinsley and R. A. Kass. The latent structure of the need satisfying properties of leisure activities. *Journal of Leisure Research,*, 11, 1979.