ISSN: 2278-4632
Vol-10 Issue-3 No.01 March 2020

Comparison of the Performance of Random Forest, Neural Network, and Support Vector Machine Models for Highway Construction Cost Prediction

[1*]*Soumya Ranjan Satapathy,* [2]*Sujit Kumar Rout*
[1*] *Professor, Dept. of Civil Engineering, NIT BBSR,*
Asst. Professor Dept. of Civil Engineering, NIT, BBSR
[1*]*soumyaranjan@thenalanda.com,sujitrout@thenalanda.com*

ABSTRACT
The final cost of construction projects is significantly impacted by inaccurate cost estimates, which also reduce profitability. Cost estimation during the conceptual stage is difficult due to the lack of knowledge. Although methodologies for cost estimating have been carefully investigated for this purpose, they are not frequently used in reality. The primary objective of this study is to evaluate how well different models perform in estimating construction project costs at the early conceptual stages of project development. Three modelling methods, including random forest, support vector machine, and artificial neural networks, are utilised in this study to anticipate the construction cost of Ethiopian highway projects using actual project data.

The results of prediction and root mean square error were then used to compare the three models. The results showed that random forest achieved superior prediction accuracy than neural network and support vector machine. Based on root mean square error, neural network and support vector machine models are 18.8% and 23.4% less accurate than the random forest cost model, respectively. It is projected that by applying a random forest regression technique in the development of a highway construction cost estimation model, a more reliable cost estimation model may be built in the early project phases. In conclusion, professionals in the Ethiopian highway construction sector are capable of making wise financial choices at the initial stages of the project's growth.

Keywords:
Cost prediction;
Highway construction project;Neural networks;
Support vector machine;Random forest.

1. Introduction

In construction practice of developing and transition countries, incompetency is a well-known fact in completing projects on time [1]. Success of construction projects is evaluated by realizing to budget, timing, and quality of work as per client's expectations Accurate cost estimation in the preliminary stage of a project is essential for decision-makers to control the overall project [2]. In addition, the importance of early estimation from the viewpoint of owners and related project teams cannot be over-emphasized [3]. Adequate estimation of construction cost is key factor in any type of construction projects. However, forecasting cost of construction projects can be considered as challenging task [4]. Moreover, Ma et al. [5] stated that construction cost estimation, which is normally labor-intensive and error-prone, is one of the most important works concerned by multi-participants during a project's life cycle. Previous studies have showed that the combination of predictive analytics and historical data can upswing cost estimation in construction projects. However, there exists a challenge in accurately estimating the cost of projects at the conceptual phase [6–8].

In order to simplify the aforementioned problems and estimate the construction project cost more accurately and rapidly, this study puts forward a method of cost estimation of construction project

based on machine learning algorithms. So, the study is going to discourse our work on predicting the cost of highway construction projects with few project features or attributes. This is a typical regression problem in which this paper aims to predict the cost of a highway project given its features. The inspiration in doing such investigation is to provide all contracting parties accurate information about the expected cost of highway projects at its early phase with minimal errors. Upon the completion of the different modelling algorithms, an evaluation of each model is conducted by comparing its accuracy.

1. Literature Review

Various estimation techniques and methods are available. With the improvements in computing capability, latest cost estimating techniques tend to use more complex approaches and a greater size of data. Machine learning algorithms as part of artificial intelligence, which allow exploring multi- and non-linear relationships between variables and final costs, have been employed in recent years [9–11]. In particular, abundant applications of support vector machines, artificial neural networks and random forest regression in the various field of civil engineering are described for prediction as well as optimization problems [12–14]. In this section, the extant literatures related to regression problems in the realm of construction are comprehensivelyreviewed.

Aiming to minimize the prediction error in conceptual estimates, Dursun and Stoy [15] adopted a multistep ahead (MSA) approach relies on the idea of using several cascaded estimations to predict future values. Based on the test outcomes obtained from 657 building projects, MSA approach significantly outperforms the prediction accuracy of linear regression (LR) and artificial neural network (ANN) techniques. Petruseva et al. [16] predicted the bidding price in construction using support vector machine (SVM). Yousefi et al [17] proposed ANN model to forecast cost and time claims in Iran construction projects. Magdum and Adamuthe [2] presented construction cost prediction models using LR and ANN and the results revealed that ANN give better prediction accuracy than statistical regression method. The accuracy of LR and SVM models in forecasting construction costs was compared in study conducted by [18]. Peško et al. [19] estimated of durations and costs of construction of urban roads using ANN and SVM. Shin [20] also developed a model using SVM to predicting the construction safety and health management cost. In recent times, random forest (RF) have been applied in various real world regression and optimization problems [21–24]. Random forest was also applied in construction world [25,26]. Kang and Ryu [14] predicted types of occupational accidents at construction sites using random forest model. In summary, estimation of cost of the construction of highways by using RF, ANNs and SVM is not present in the literature. This study is, therefore, aimed to make comparison of the performance of the three models in forecasting the cost of highway projects.

2. Methodology

In the process of dealing with a regression problem described in this study, k-fold cross validation and Root Mean Squared Error (RMSE) metric is employed to run and validate the modelling process and make a comparative assessment respectively. RMSE is the most importantcriterion for fit if the main purpose of the model is a prediction [6]. In particular, the methodology followed in

ISSN: 2278-4632
Vol-10 Issue-3 No.01 March 2020

this study is: (a) each method trained using 3-fold cross-validation and
(b) final RMSE is computed based on the average results of all training steps. Three model algorithms such as SVM, ANN and RF are planned to be performed using Python programming with Scikit-Learn library packages. All study outcomes are prepared and presented using the various Python programing packages.

4. Variables Analysis

Data Collection and Description

The historical project data was compiled by authors from the Ethiopian Road Authority (ERA) management system software. The highway projects which have been started and completed between 2006 and 2018 are considered in the process of developing the historical data base. The project costs data set has 8 variables (features) to predict the cost in which 4 numerical variables and 4 categorical variables. The variables include project length, number of bridges, inflation rate, project scope, terrain type, project type, contract duration and project location. There are 74 project cases considered in this study. The project data are recorded and the model dataset are compiled based on the above-mentioned input variables. The description of the dataset is summarized in Fig.

	Project_Length	No_of_Bridges	Inflation_Rate	Project_Scope	\
count	74.000000	74.000000	74.000000	74.000000	
mean	68.905405	4.391892	15.202135	1.527027	
std	35.569407	5.411347	4.560355	0.831288	
min	10.000000	0.000000	8.225000	1.000000	
25%	44.250000	0.000000	12.425000	1.000000	
50%	63.500000	2.500000	16.180000	1.000000	
75%	91.000000	6.750000	18.200000	2.000000	
max	180.000000	21.000000	25.250000	3.000000	

	Terrain_Type	Project_Type	Contract_Duration	Project_Location	\
count	74.000000	74.000000	74.000000	74.000000	
mean	1.762859	2.171723	951.094595	2.608108	
std	0.511248	1.104458	236.731058	1.488012	
min	1.000000	1.000000	90.000000	1.000000	
25%	1.465775	1.000000	910.000000	1.000000	
50%	1.679250	2.000000	1065.000000	2.000000	
75%	2.000000	2.170625	1095.000000	4.000000	
max	2.700000	4.000000	1280.000000	5.000000	

	Project_Cost
count	7.400000e+01
mean	5.187602e+08
std	3.727179e+08
min	7.011855e+05
25%	1.899287e+08
50%	4.903572e+08
75%	8.078543e+08
max	1.533714e+09

ISSN: 2278-4632
Vol-10 Issue-3 No.01 March 2020

Data Preprocessing

Fig. 1. Description of project dataset.

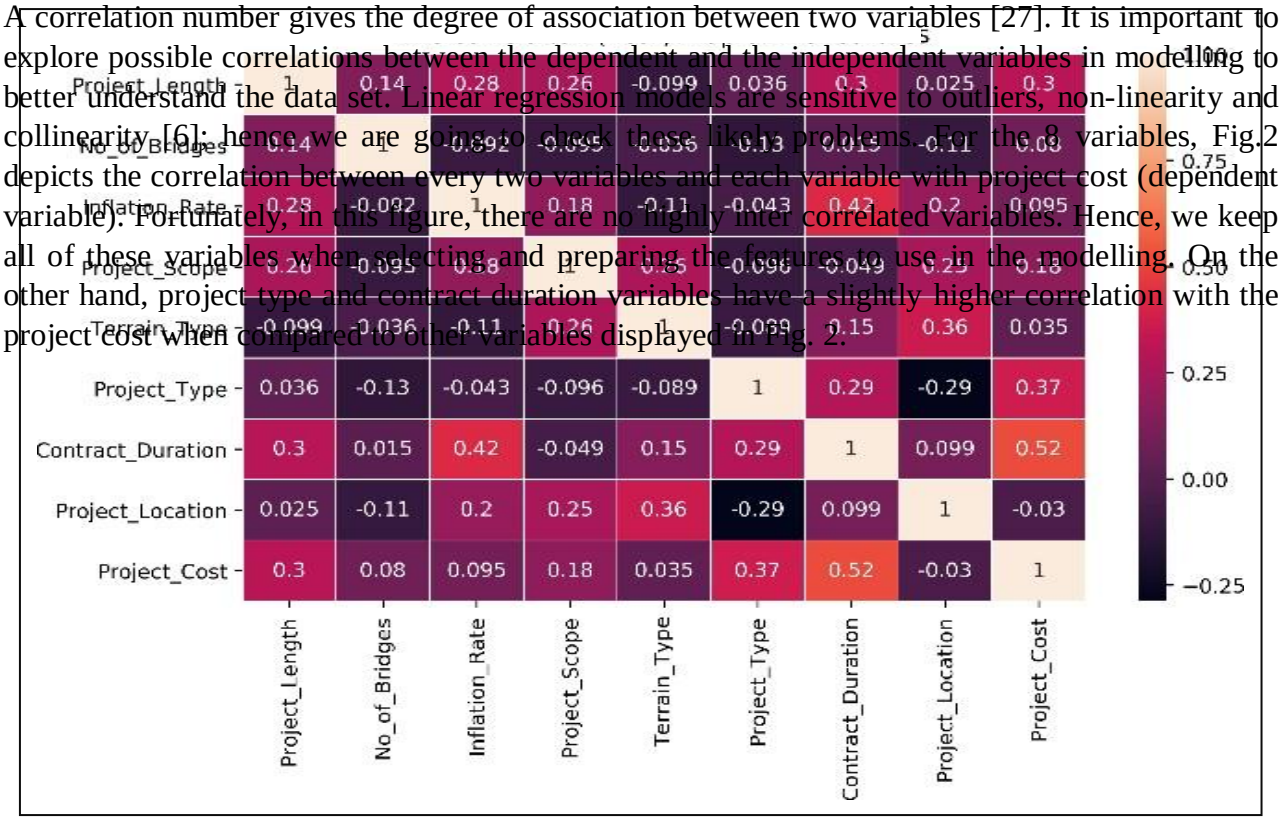
Data preprocessing is important to produce outputs that can be smoothly utilized as inputs in data modeling by transforming the raw data. At this stage, the categorical variables are converted to numerical for the sake of model simplicity. In addition, the correlation matrix is generated to investigate the possible relationship among model input variables to minimize the likely impact on the prediction outcome.

Convert categorical variable to numerical variable

Several machine learning based modelling algorithms need numbers as inputs, it requires to be coded as numbers in some way. Accordingly, the four features are coded with numbers before the modelling process get started.

Correlation matrix or heatmap

A correlation number gives the degree of association between two variables [27]. It is important to explore possible correlations between the dependent and the independent variables in modelling to better understand the data set. Linear regression models are sensitive to outliers, non-linearity and collinearity [6]; hence we are going to check these likely problems. For the 8 variables, Fig.2 depicts the correlation between every two variables and each variable with project cost (dependent variable). Fortunately, in this figure, there are no highly inter correlated variables. Hence, we keep all of these variables when selecting and preparing the features to use in the modelling. On the other hand, project type and contract duration variables have a slightly higher correlation with the project cost when compared to other variables displayed in Fig. 2.



ISSN: 2278-4632
Vol-10 Issue-3 No.01 March 2020

Fig. 2. Correlation of input variables with project cost.

Log transformation of the dependent variable

As recommended by the specialists, log transformation on the dependent variable i.e. project costis applied. Log transform can transform data into ones that are symmetric and skewed. Relatively, it moves smaller values farther apart while it moves big values closer together (see

Fig. 3). This is the most imperative feature of log transformation [28]. Moreover, it is also easier to describe the relationship between variables when it's approximately linear. Generally, it workswell in modeling to treat the project cost data up to a high amount.

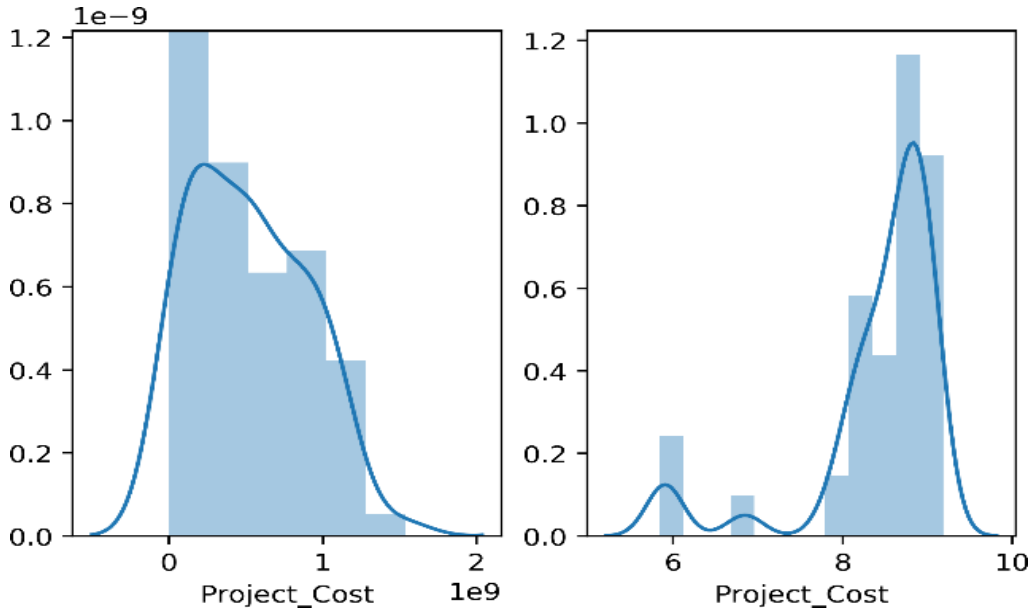


Fig. 3. Histogram of project cost with/ without log transformation.

Feature scaling

Standardizing the data is another important step in the preprocessing phase because this will be helpful for all the models. The scaling was done independently for the training and the testing sets as cross-validation is employing in this study.

5. Performance evaluation of models

This section presents the evaluation of different scikit-learn modeling algorithms. The final step is to evaluate the performance of each modeling algorithm. This step is particularly important to compare how well different algorithms perform on a certain dataset. In this study, RMSE evaluation metric which is the square root of the mean of the squared errors [28] is used as RMSE amplifies and severely punishes large errors [10] and its equation is written as follows:

ISSN: 2278-4632
Vol-10 Issue-3 No.01 March 2020

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \tilde{y}_j)^2} \quad (1)$$

Where y_j stands for log (Project_Cost$_j$) and $\tilde{y}_j$ stands for log (predicted Project_Cost$_j$). Fortunately, it is not required to perform this calculation manually. The Scikit-Learn library comes with pre-built functions that can be used to find out RMSE value. Finally, the RMSE results are utilized to make a comparative analysis as it shows the prediction accuracy of all the models.

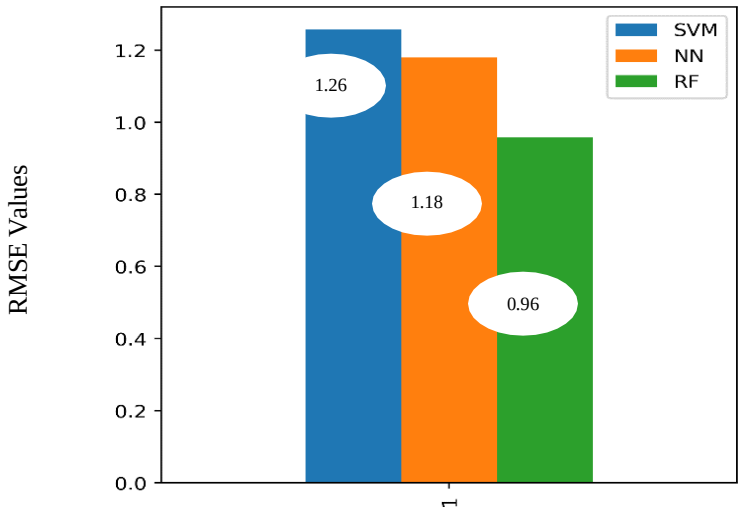
6. Results and discussions

The ANN, SVM and RF models were developed using the test dataset for predicting in the final cost of highway projects. The performance of the three models were evaluated using the model accuracy measures tabulated in Table 1.

Table 1. Model performance results

Models	Average RMSE
SVN model	1.2569
ANN model	1.1802
RF model	0.9579

For the three models, RF helps practitioners or researchers acquire the most accurate prediction outcome in this cost data set with smaller error value. Conversely, SVM provides the worst result compared with ANN algorithms because lower values of RMSE indicate better fit. Based on RMSE values, the RF cost model provides 18.8% and 23.4% more accurate result than ANN and SVM models respectively.



ISSN: 2278-4632
Vol-10 Issue-3 No.01 March 2020

Prediction Models

Fig. 4. Comparison of model performances based on RMSE**.**

The RF model predicted the cost of highway projects with RMSE value of 0.96 i.e., the difference between predicted and actual cost values were insignificant. Fig. 5 clearly portrays that the predicted project cost values were in strong coherence with those of actually collected cost values. This justifies that the RF model was able to generate the predicted cost results accurately when it compared to ANN and SVM models.

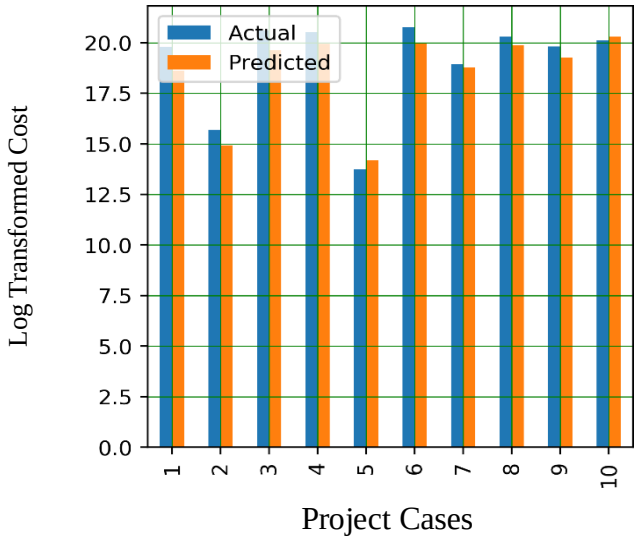


Fig. 5. Comparison of FR actual and predicted values

7. Conclusions

.The main objective of this study was to develop models for predicting the cost of highway projects and make a comparative assessment based on their accuracy using RMSE results. Allthe necessary computations including model developments were performed using different Scikit-Learn library packages in the Python programing. In this study, SVM, NN and RFalgorithms were employed to forecast highway project costs. The results clearly revealed that RFhas more accuracy in prediction with less error value when compared with NN and SVM. It can be generalized that the prediction done with RF portrays a strong degree of coherency with actually collected cost data of highway project against NN and SVM. So, this study will be helpful the contracting parties in the highway construction industry and the future works. A mobile app or simple desktop package can be created by storing the predicted data in the databases so that the contracting parties would really have a brief information and would safely invest the money on the proposed project.

References

[1] Bayram S. Duration prediction models for construction projects: In terms of cost or physical characteristics? KSCE J Civ Eng 2017;21:2049–60. doi:10.1007/s12205-016-0691-2.

[2] Magdum SK, Adamuthe AC. Construction Cost Prediction Using Neural Networks.

ISSN: 2278-4632
Vol-10 Issue-3 No.01 March 2020

ICTACT JSoft Comput 2018;8:1549–56. doi:10.21917/ijsc.2017.0216.

[3] Kim S, Shim JH. Combining case-based reasoning with genetic algorithm optimization for preliminary cost estimation in construction industry. Can J Civ Eng 2014;41:65–73. doi:dx.doi.org/10.1139/cjce-2013-0223.

[4] Arage SS, Dharwadkar N V. Cost Estimation of Civil Construction Projects using Machine Learning Paradigm. Int Conf I-SMAC (IoT Soc Mobile, Anal Cloud) (I-SMAC 2017), 2017, p. 594–9.

[5] Ma Z, Liu Z, Wei Z. Formalized Representation of Specifications for Construction Cost Estimation by Using Ontology. Comput Civ Infrastruct Eng 2016;31:4–17. doi:10.1111/mice.12175.

[6] Bouras CB-T. Regression models to house price prediction. 2018.

[7] Chau AD, Moynihan GP, Vereen S. Design of a Conceptual Cost Estimation Decision Support System for Public University Construction. Constr Res Congr 2018, Reston, VA: American Societyof Civil Engineers; 2018, p. 629–39. doi:10.1061/9780784481295.063.

[8] Mayer M, Bourassa SC, Hoesli M, Scognamiglio D. Estimation and updating methods for hedonic valuation. J Eur Real Estate Res 2019;12:134–50. doi:10.1108/JERER-08-2018-0035.

[9] Matel E, Vahdatikhaki F, Hosseinyalamdary S, Evers T, Voordijk H. An artificial neural network approach for cost estimation of engineering services. Int J Constr Manag 2019;0:1–14. doi:10.1080/15623599.2019.1692400.

[10] Cao Y, Ashuri B, Baek M. Prediction of Unit Price Bids of Resurfacing Highway Projects through Ensemble Machine Learning. J Comput Civ Eng 2018;32:04018043. doi:10.1061/(ASCE)CP.1943-5487.0000788.

[11] Poh CQX, Ubeynarayana CU, Goh YM. Safety leading indicators for construction sites: A machine learning approach. Autom Constr 2018;93:375–86. doi:10.1016/j.autcon.2018.03.022.

[12] Golizadeh H, Banihashemi S, Sadeghifam AN, Preece C. Automated estimation of completion time for dam projects. Int J Constr Manag 2017;17:197–209. doi:10.1080/15623599.2016.1192249.

[13] Rafiei MH, Adeli H. Novel Machine-Learning Model for Estimating Construction Costs Considering Economic Variables and Indexes. J Constr Eng Manag 2018;144:04018106. doi:10.1061/(ASCE)CO.1943-7862.0001570.

[14] Kang K, Ryu H. Predicting types of occupational accidents at construction sites in Korea using random forest model. Saf Sci 2019;120:226–36. doi:10.1016/j.ssci.2019.06.034.

[15] Dursun O, Stoy C. Conceptual estimation of construction costs using the multistep ahead approach.J Constr Eng Manag 2016;142:1–10. doi:10.1061/(ASCE)CO.1943-7862.0001150.

[16] Petruseva S, Sherrod P, Pancovska VZ, Petrovski A. Predicting Bidding Price in Constructionusing Support Vector Machine. TEM J J 2016;5:143–51. doi:10.18421/TEM52-04.

[17] Yousefi V, Yakhchali SH, Khanzadi M, Mehrabanfar E, Šaparauskas J. Proposing a neural network model to predict time and cost claims in construction projects. J Civ Eng Manag 2016;22:967–78. doi:10.3846/13923730.2016.1205510.

ISSN: 2278-4632
Vol-10 Issue-3 No.01 March 2020

[18] Petruseva S, Zileska-pancovska V, Vahida Ž, Vejzović AB-. Construction Costs Forecasting : Comparison of the Accuracy of Linear Regression and Support Vector Machine Models. Tech Gaz 2017;24:1431–8.

[19] Peško I, Mučenski V, Šešlija M, Radović N, Vujkov A, Bibić D, et al. Estimation of costs and durations of construction of urban roads using ANN and SVM. Complexity 2017;2017:1–13. doi:10.1155/2017/2450370.

[20] Shin SW. Construction Safety and Health Management Cost Prediction Model using Support Vector Machine. J Korean Soc Saf 2017;32:115–20.doi:https://doi.org/10.14346/JKOSOS.2017.32.1.115.

[21] Neloy AA, Haque HMS, Ul Islam MM. Ensemble Learning Based Rental Apartment Price Prediction Model by Categorical Features Factoring. Proc 2019 11th Int Conf Mach Learn Comput
- ICMLC '19, New York, New York, USA, China: ACM Press; 2019, p. 350–6. doi:10.1145/3318299.3318377.

[22] Angamuthu Chinnathambi R, Mukherjee A, Campion M, Salehfar H, Hansen T, Lin J, et al. A Multi-Stage Price Forecasting Model for Day-Ahead Electricity Markets. Forecasting 2018;1:26– 46. doi:10.3390/forecast1010003.

[23] BaniMustafa A. Predicting Software Effort Estimation Using Machine Learning Techniques. 2018 8th Int Conf Comput Sci Inf Technol, IEEE; 2018, p. 249–56. doi:10.1109/CSIT.2018.8486222.

[24] Torres-Barrán A, Alonso Á, Dorronsoro JR. Regression tree ensembles for wind energy and solar radiation prediction. Neurocomputing 2019;326–327:151–60. doi:10.1016/j.neucom.2017.05.104.

[25] Alaka HA. 'Big data analytics' for construction firms insolvency prediction models. The Universityof the West of England, 2017.

[26] Bai S, Li M, Kong R, Han S, Li H, Qin L. Data mining approach to construction productivity prediction for cutter suction dredgers. Autom Constr 2019;105:102833. doi:10.1016/j.autcon.2019.102833.

[27] Shinde N, Gawande K. Valuation of House Prices Using Predictive Techniques. 2018.

[28] Lu S, Li Z, Qin Z, Yang X, Goh RSM. A hybrid regression technique for house prices prediction. IEEE Int Conf Ind Eng Eng Manag, 2018. doi:10.1109/IEEM.2017.8289904.