A MODEL FOR CONSTRUCTION SCHEDULE AND COST PREDICTION USING REGULARIZED GRADIENT BOOSTED REGRESSION TREE ALGORITHM

By

S. M. Tazim Ahmed

A thesis
submitted to the
Department of Industrial and Production Engineering,
Bangladesh University of Engineering and Technology
in partial fulfillment of the requirements
for the Degree
of
MASTER OF SCIENCE
in
Industrial and Production Engineering

Department of Industrial and Production Engineering (IPE)
Bangladesh University of Engineering and Technology (BUET), Bangladesh

September, 2020

CERTIFICATE OF APPROVAL

The thesis titled **"A MODEL FOR CONSTRUCTION SCHEDULE AND COST PREDICTION USING REGULARIZED GRADIENT BOOSTED REGRESSION TREE ALGORITHM"** submitted by S. M. Tazim Ahmed, Student ID.: 0416082022, Session: April, 2016, has been accepted as satisfactory in partial fulfillment of the requirements for the degree of Master of Science in Industrial and Production Engineering on 13 September, 2020.

BOARD OF EXAMINERS

Dr. Ferdous Sarwar Associate Professor Department of IPE, BUET, Dhaka	Chairman (Supervisor)
Dr. Nikhil Ranjan Dhar Professor and Head Department of IPE, BUET, Dhaka	Member (Ex-officio)
Dr. Syed Mithun Ali Associate Professor Department of IPE, BUET, Dhaka	Member
Shuva Ghosh Dr. Shuva Ghosh Associate Professor Department of IPE, BUET, Dhaka	Member
Dr. Tanvir Ahmed Professor Department of Civil Engineering, BUET, Dhaka	Member (External)

CANDIDATE'S DECLARATION

It is hereby declared that this thesis or any part of it has not been submitted elsewhere for the award of any degree or diploma.

Tazim

S. M. Tazim Ahmed

Student ID: 0416082022

ABSTRACT

Accurate prediction of construction schedule and cost plays critical role to project success. Many quantitative and associative models have been developed for more accurate prediction. However, these models often lack robustness due to bias and variance. Ensemble type of machine learning algorithm can perform well for prediction by balancing bias and variance.

This study aims to develop construction schedule and cost prediction model using one of the recent ensemble machine learning algorithms named Gradient Boosted Regression Tree (GBRT). Data were obtained from 69 construction projects of Dhaka city of Bangladesh. These projects were categorized as low rise, medium rise and high rise buildings according to the number of floors. One-way ANOVA F-test has been applied to select the statistically significant features. Finally, the regularized GBRT has been applied to develop the construction schedule and cost prediction models. Performances of regularized GBRT models were compared to Support Vector Regression (SVR) and Multiple Linear Regression (MLR) models. Mean absolute percentage error (MAPE) and mean squared error (MSE) were used as performance metrics. One-way ANOVA feature selection method reveals that location, land size, floor height, floor area, number of basement, workforce level and number of floor had significant impact on schedule and cost prediction model for low rise buildings. For medium and high rise buildings, land size, floor area, number of basement, workforce level and number of floor are the most significant features. The results show that regularized GBRT models have lower MAPEs and MSEs than SVR and MLR models. Therefore, regularized GBRT models have performed better than SVR and MLR models in construction schedule and cost prediction for low, medium and high rise buildings.

ACKNOWLEDGEMENT

First, the author would like to express his deepest gratefulness to the most benevolent and Almighty God, because without His grace and mercy it was quite impossible to complete this thesis. Then the author also would like to extend thanks to his family for their continuous inspiration, sacrifice and support to complete the thesis successfully.

The author expresses sincere respect and gratitude to his thesis supervisor Dr. Ferdous Sarwar, Associate Professor, Department of Industrial and Production Engineering, BUET, Dhaka-1000, under whose supervision this thesis has been carried out. His guidance, valuable suggestions and inspirations throughout this work made the study possible.

The author also expresses his sincere gratitude to the board members- Dr. Nikhil Ranjan Dhar, Professor and Head, Department of IPE, BUET, Dr. Syed Mithun Ali, Associate Professor, Department of IPE, BUET, Dr. Shuva Ghosh, Assistant Professor, Department of IPE, BUET, and Dr. Tanvir Ahmed, Professor, Department of Civil Engineering, BUET, for their constructive remarks and evaluation of this research.

Finally, the author wishes to express deepest sense of gratitude to all of his colleagues and friends for their kind co-operations and inspirations provided during this research work.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

GBRT	:	Gradient Boosted Regression Tree
SVR	:	Support Vector Regression
ANOVA	:	Analysis of Variance
CBR	:	Case Based Reasoning
CMQ	:	Construction Material Quality
EVM	:	Earned Value Management
MAE	:	Mean Absolute Error
MAPE	:	Mean Absolute Percentage Error
MSE	:	Mean Squared Error

CHAPTER 1: INTRODUCTION

1.1 Background of the Study

Prediction and estimation of construction project schedule and cost are very crucial steps of construction management for the project owners, contractors, and estimators because construction schedule and cost tend to fluctuate due to the variation of different parameters **[Bayram, 2017; Y.-J. Kim et al., 2019]**. As a result, it affects the budget estimation as well as pricing system. Therefore, construction project success is largely dependent on the early and accurate prediction of construction duration and cost **[Shayboun & Koch, 2019]**. However, it becomes very difficult for the estimators and project managers to predict and estimate the construction duration and cost accurately at the preliminary planning stage because the drawings and other documents are still incomplete. Meanwhile, business success largely depends on the project success. Management and mitigation of inherent risks also require early and accurate prediction of construction schedule and cost **[Sullivan et al., 2017]**. Therefore, development of some expert systems is required to perform these jobs of prediction as accurately as possible.

Over the past few decades, researchers and construction engineers have explored the importance of model development for the prediction of construction duration and cost. Many quantitative and associative models have been developed for more accurate prediction. Most of these models face difficulties for big amount of data with large number of variables **[Boon et al., 2019]**. In this case, different supervised machine learning algorithms are being applied to predict construction duration and costs **[Arage & Dharwadkar, 2017]**. Many construction firms want to use these machine learning algorithms to predict schedule and costs for improving the business performance. However, they may fail due to lack of selecting important features and appropriate algorithms for model development.

Many researchers have applied different machine learning algorithms and statistical techniques to develop construction duration and cost prediction models. Elfahham **[2019]** used some machine learning techniques such as neural network, linear regression and also auto regressive time series method to estimate and predict the construction cost index in Egypt. Thomas and Thomas & Thomas **[2016]** also developed multiple regression model for construction cost and duration prediction. Lin et al. **[2019]** came about an intelligent prediction model of the construction cost using support vector machine (SVM) algorithm for substation projects. They optimized the hyper parameters of SVM using particle swarm optimization algorithm. However, to the best of our knowledge, the literature of construction project management still lacks the application of ensemble type of machine learning algorithms for developing construction duration and cost prediction models. Gradient boosted regression tree is one of the popular recent ensemble type machine learning algorithms.

1.2 Objectives with Specific Aims

The specific objectives of this research are-

a) To identify the most significant features for developing construction schedule and cost prediction models for low rise, medium rise and high rise buildings.

b) To select the best values of hyper-parameters of GBRT models using random search method.

c) To develop construction schedule and cost prediction models for each category of building using regularized Gradient Boosted Regression Tree (GBRT) ensemble machine learning algorithm.

d) To compare the performance of the regularized GBRT models with Support Vector Regression (SVR) and Multiple Linear Regression (MLR) models.

1.3 Outline of the Methodology

The methodology to achieve the objectives of the research is outlined below:

a) Different features or criteria for construction schedule and cost prediction of different categories of building have been identified through reviewing previous literature and getting feedback from the industry experts.

b) A framework has been developed to predict construction schedule and cost using Gradient Boosted Regression Tree (GBRT) algorithm for low, medium and high rise buildings.

c) To validate the framework, data for those criteria identified in (a) have been collected and data preprocessing with exploratory data analysis has been performed.

d) Statistically significant features or criteria have been selected using one-way ANOVA method.

e) The criteria identified in (d) have been used as the input features of the data set of construction schedule and cost prediction model.

f) Construction schedule and cost prediction models have been developed using regularized Gradient Boosted Regression Tree (GBRT) algorithm. Regularization of the GBRT has been performed using Random Search method for reducing over fitting problem and getting best results.

g) Finally, the performance of the GBRT models have been compared with Support Vector Regression (SVR) and Multiple Linear Regression (MLR) models.

1.4 Contributions of the Present Study

This research proposes construction project schedule and cost prediction models for low, medium and high rise buildings based on regularized Gradient Boosted Regression Tree (GBRT) algorithm. Different machine learning algorithms like Artificial Neural Network (ANN), Multiple Linear Regression (MLR), Support Vector Regression (SVR) etc. have been applied for predicting construction schedule and cost. However, sometimes these models lack robustness due to bias-variance trade-off. Ensemble type of machine learning algorithm can help to solve this issue. Gradient Boosted Regression Tree (GBRT) is one of the recent ensemble type of machine learning algorithms and to the best of our knowledge, this algorithm has not yet been applied in construction schedule and cost prediction. This research contributes to the construction project management literature by developing GBRT based prediction models.

In this research, one-way ANOVA F-test has been applied for selecting the most significant features for developing the prediction model. Integration of one-way ANOVA in the GBRT model is also a new addition in developing construction schedule and cost prediction.

Selection of best hyper-parameter values of GBRT models for best performance using Random Search method is another contribution of the study on which a very few work has been done. Finally, the performances of the developed models have been compared with the Support Vector Machine and Multiple Linear Regression for validation.

1.4 Organization of the Thesis

Chapter 1 presents the background of the research work, objectives with specific aims, contributions of the study and outline of the methodologies.

Chapter 2 is comprised of the relevant literature on construction schedule and cost prediction, use of machine learning algorithms in schedule and cost prediction and application of Gradient Boosted Regression Tree (GBRT).

Chapter 3 provides the theoretical background of one-way ANOVA, ensemble machine learning algorithms, Gradient Boosted Regression Tree (GBRT) algorithm, performance metrics, and methods of regularization.

Chapter 4 explains of the framework development for construction schedule and cost prediction. This chapter describes the development of the proposed framework using one-way ANOVA, GBRT and random search regularization method.

Chapter 5 is comprised of the data collection process and implementation of the proposed framework on the data to predict construction schedule and cost prediction for low, medium and high rise buildings.

Chapter 6 contains the possible interpretation on the results obtained from the implementation of the proposed model. This chapter also includes the detailed discussion on the findings.

Finally, **Chapter 7** contains the conclusion of this work and the scope of future research associated with this work.

CHAPTER 2: LITERATURE REVIEW

This section discusses the previous works on construction schedule and cost prediction models, application of machine learning algorithms in construction schedule and cost prediction, and gradient boosted regression tree algorithm (GBRT).

2.1 Construction Schedule and Cost Prediction Models

During the past few decades, researchers and construction engineers have explored different models for construction schedule and cost prediction.

Al-Momani **[1996]** proposed an analytical model for estimating public school building construction costs. Observations of 125 school projects were performed in Jordan from 1984-1994. They found that when a project was finished, the real cost exceeded the initial contract price by 30%, while the change orders resulted in a cost overrun of 8.3%. Capital spending on school programs over the next 5 years was expected at JD 695,9 million.

The performance of three cost estimate models was analyzed by Kim et al. **[2004]**. The reviews were focused on the 530 historic cost data from multiple regression analysis (MRA), neural networks (NNS) and case-based reasoning (CBR). The best NN estimating model provided more reliable estimates than either the MRA or CBR estimates. However, with respect to long-term use, available knowledge from the results and time versus precision compromise, the CBR estimating model worked better than the NN estimating model.

In the context of road construction projects in the States of Florida, USA, Shr & Chen **[2006]** created a structure for the concept of building expenses and time. The model suggested providing the State Highway Agencies (SHAs) and contractors with more control and an appreciation of the time value of road construction projects. This model is not however, sufficient for projects with a high degree of change orders.

Abu Hammad et al. **[2008]** developed a probabilistic model for the construction projects to predict the risk effect on construction cost and time. This time and cost prediction model was developed based on historical data. This model was based on historic data to estimate project cost and duration. On the actual data from 140 construction projects in Jordan, they employed linear regression and multiple linear regression. However, this model is limited to numerical variable and linear data.

Many researchers have widely used case-based reasoning (CBR) technique for developing duration and cost prediction models **[Koo et al., 2018]**. Koo et al., **[2010]** developed a case-based reasoning (CBR)-based hybrid model with which to predict the construction duration and cost of a project in its early stage. One hundred and one cases among multi-family housing projects that were completed between 2000 and 2005 were used. The CBR-based hybrid model developed in this study was the result of integrating the advantages of (*i*) prediction methodologies, such as case-based reasoning, multiple regression analysis, and artificial neural networks, (*ii*) the optimization process using a genetic algorithm, and (*iii*) the probability distribution and the analysis process of outlier using Monte-Carlo simulation.

Jin et al. **[2012]** built an enhanced CBR model that integrates multiple regression analysis techniques to early predict construction costs. This research was conducted on 41 company premises and 99 international housing projects. The result showed a 17.23% and 4.39% improved projected performance of the updated CBR model for business facilities and multi-family homes compared with the current CBR model respectively. The proposed updated CBR MRA model was supposed to be helpful in the initial project process in estimating construction costs.

Ahn et al. **[2017]** also applied the CBR technique for the early prediction of the construction costs. For accurate and reliable cost estimation, quality source data are required and the existing similarity measures have limitations in taking the covariance among attributes into consideration. In their works, they mentioned this challenge and to deal with this challenging issues, they examined the weighted Mahalanobis distance based similarity measure applied to CBR cost estimation.

Case based reasoning method was revised by Jin et al. **[2014]** considering the deviation of categorical and numerical variables using regression analysis. They applied the revised method to multi-housing projects for cost prediction at the early stage of construction. These studies, however, were limited in considering the deviation of numerical variables while the majority of variables available in the early stage was more categorical (e.g., structural system or underground condition) than numerical (e.g., gross floor area).

Soto & Adey **[2015]** have used case-based reasoning (CBR), to estimate building projects capital. They used the nearest neighbor technique to evaluate the similarity for the retrieval process to estimate the construction materials quantities (CMQs) in structural concrete. Two types of distances, i.e. 1) the City-block distance and 2) the Euclidean distance, and four different types of weights, based on regression analysis and feature counting, to account for the relative importance of the different parameters, were investigated. The four different types of weights used were 1) the adjusted unstandardized coefficients from the regression models, 2) the unadjusted unstandardized coefficients from the regression models, 3) the standardized coefficients from the regression models, and 4) equal weights.

Mackova et al. **[2017]** developed prediction model based on computer experiment to estimate the construction duration of residential buildings in Slovak republic. They considered the gross floor area, number of stories and floor area of one storey as variable inputs to predict the time duration for construction while taking into account the intensity of the deployment of labor resources. They used multiple linear regression analysis to develop the construction duration prediction model.

Chen **[2018]** examined the longitudinal relationships between completed project time and project variables' performance prior to the construction phase and to model those useful relationships to predict completed project duration. The study confirmed the significance of scope, team, communication, risk, and innovation performance prior to the construction phase and revealed reasonable estimation accuracy and a relatively small difference in prediction rates between in-sample (88.13%) and out-of-sample (85.49%) data.

Apart from these, many time series techniques have been applied by researchers for construction duration and cost prediction. Niu & Hua **[2016]** built cost index prediction model using ARIMA and exponential smoothing technique for power transmission and transformation project. Ng et al. **[2004]** developed an integrated regression analysis (RA) and time series (TS) model for predicting construction tender price index. The performance of integrated RA-TS model has been compared to individual regression analysis (RA) and time series (TS) model.

In conjunction with the well-established exponential smoothing technique, Khamooshi & Abdi **[2017]** developed a model construction duration prediction with EDI (earned duration index) method. The trial compares a number of model models with the use of different projects in various phases of completion and provides a comparative review of their results.

Batselier & Vanhoucke **[2017]** performed the similar type of study by extending the traditional earned value management (EVM) and earned schedule (ES) methodology integrating exponential smoothing method with these approaches for construction time and cost forecasting. This study results in an extension of the known EVM and earned schedule (ES) cost and time forecasting formulas. Mao & Xiao **[2019]** also applied time series forecasting technique for construction cost index prediction. First, time series data were mapped into a network by visibility graph. Then, the link prediction method was adopted to calculate the similarity index.

Recently, different machine learning algorithms have been applied widely for developing more improved duration and cost prediction model.

2.2 Application of Machine Learning Algorithms in Construction Schedule and Cost Prediction

Machine learning models rely on historical data and statistical inference **[Niu et al., 2017]**. Many researchers have developed construction schedule and cost prediction models using supervised machine learning algorithms. Artificial neural network is one of the widely used approach to build the construction cost and duration prediction models **[Bayram et al., 2016; Waziri et al., 2017]**.

Mensah et al. **[2016]** developed a consistent model of early design cost prediction of road construction and the duration of the project using neural network approach. Data for 22 completed bituminous surfaced road projects from the Department of Feeder Roads (rural road agency) were collected and analyzed using the principal component analysis (PCA) and ANN techniques. The data collected were final payment certificates which contained payment bill of quantities (BOQ) of work items executed for the selected completed road projects. The ANN was then used to develop the network using the identified significant quantities as input variables and the actual durations as output variables.

Naik & Radhika **[2015]** developed different artificial neural network (ANN) models for the estimation of cost and duration for highway road construction projects. They collected data from the completed projects and after normalizing, they used the data as inputs and targets for developing ANN models.

Petroutsatou et al. **[2012]** built early cost prediction model for road tunnel construction projects using neural network approach. First, the basic parameters (geological, geometrical, and work quantities-related) affecting temporary and permanent support and final construction cost were determined. Appropriate price lists were then applied to calculate the costs; subsequently, cost-estimating models were developed using two types of neural networks: (1) the multilayer feedforward network; and (2) the general regression neural network. Finally, these models were compared against real quantities and costs for accuracy and robustness.

Leu & Liu **[2016]** used back-propagation neural network (BP-NN) for predicting the construction duration during the initial stage. Here, they applied principle component analysis (PCA) for selecting the key features as input data for developing the duration prediction model. Principal component analysis was applied to the database to identify key factors to serve as input data for a back-propagation neural network (BP-NN) that was used to estimate the project duration. Three prediction models were identified and developed separately based on the total cost for large, medium, and small construction projects.

Many researchers have applied another popular machine learning algorithm namely support vector machine (SVM). Cheng et al. **[2010]** constructed an evolutionary estimate at completion method to estimate final construction project costs. They fused two artificial intelligence methods, namely the fast messy genetic algorithm (fmGA) and support vector machine (SVM), to create an evolutionary support vector machine inference model (ESIM).

Wang et al. **[2012]** applied artificial neural network (ANN) and support vector machine (SVM) algorithms in order to predict project cost and plan success, using early plan status as model inputs. Data were collected from 92 building projects. A comparative analysis was performed by Kim et al. **[2013]** to compare the accuracy of three estimating techniques regression analysis (RA), neural network (NN) and support vector machine techniques (SVM) by performing estimations of construction costs.

Petroutsatou et al. **[2012]** also used support vector machine for construction cost forecasting. Some researchers optimized the value of hyper-parameters of support vector machine algorithms in many models. Cheng & Hoang **[2014]** utilized least squares support vector machine (LS-SVM), machine learning based interval estimation (MLIE), and differential evolution (DE) to establish a novel model for predicting construction project cost.

Project schedule and cost forecasting models were presented by Wauters & Vanhoucke **[2014]** that compared the performance of support vector regression model with the best performing earned value and earned schedule methods. The parameters of the SVM were tuned using cross validation and grid search procedure.

Yi et al. **[2018]** applied particle swarm optimization algorithm for selecting the best values of hyper-parameters of least square support vector machine (LSSVM) for predicting the construction cost of transmission line projects. Principal component analysis (PCA) was used to reduce the dimension of indexes and particle swarm optimization (PSO) was innovatively introduced to optimize the parameters of LSSVM model to obtain the optimal parameters. The obtained principal component data were imported into empirical parameter LSSVM prediction model and the optimized parameter PSO-LSSVM prediction model, respectively, for modeling and prediction.

Support vector machine has also been integrated with Bromilow TCM for road structures construction cost prediction by Petrusheva et al. **[2019]**. Five hybrid models have been built for comparison purposes: SVM-Bromilow TCM, LR-Bromilow TCM, RBFNN-Bromilow TCM, MLP-Bromilow TCM and GRNN-Bromilow TCM, combining Bromilow TCM with SVM, LR (linear regression), RBFNN (radial basis function neural network), MLP (Multilayer perceptron) and GRNN (general regression neural network), respectively.

Luu & Kim **[2009]** applied neural network approach to estimate the construction costs of apartment projects in Vietnam. Ninety-one questionnaires were collected to identify input variables. Fourteen data sets of completed apartment projects were obtained and processed for training and generalizing the neural network(NN).

Magdum & Adamuthe **[2017]** also developed neural network and multilayer perceptron based model for construction cost prediction. Different models of NN and MLP are developed with varying hidden layer size and hidden nodes. Four artificial neural network models and twelve multilayer perceptron models are compared.

Juszczyk **[2019]** developed early cost prediction model for bridge construction using support vector machine algorithm. Different regression models have also been developed by many researchers. Mahamid **[2019]** built multiple linear regression model for the early prediction of road construction duration. El-Dash et al. **[2019]** also developed duration prediction model based on regression analysis. A small number of construction duration and cost prediction models have been developed based on ensemble machine learning algorithms **[Choi et al., 2018; Juszczyk & Leśniak, 2019; Ugur et al., 2019]**.

2.3 Application of Gradient Boosted Regression Tree Algorithm

In recent years, Gradient Boosted Regression tree (GBRT) has become very popular and it has widely been applied in different areas such as energy consumption forecasting, solar power forecasting, and web search ranking **[Mohan et al., 2011; Persson et al., 2017]**. Gradient

Boosted Regression Tree is an ensemble type of machine learning algorithm which can be used for classification and regression **[Ponomareva et al., 2017]**.

Williams & Gong **[2014]** used ensemble machine learning algorithm to predict the cost overrun of projects. Features were selected based on correlation and regression analysis. The stacking ensemble model had an average accuracy of 43.72% for five model runs. The model performed best in predicting projects completed with large cost overruns and projects near the original low bid amount. It was found that a stacking model that used only numerical data produced predictions with lower precision and recall.

Hu et al., **[2015]** used ensemble machine learning algorithm decision tree based on bagging to develop a model for outsourced software project risk prediction. Comparative analysis with T-test on 60 different risk prediction models using 327 outsourced software project samples suggested that the ideal model has been a homogeneous ensemble model of decision trees (DT) based on bagging. Interestingly, DT underperformed Support Vector Machine (SVM) in accuracy (i.e., assuming equal misclassification cost), but outperformed in cost-sensitive analysis under the proposed framework.

Torres-Barrán et al. **[2019]** developed models using GBRT algorithm for predicting wind energy and solar radiation. Besides a complete exploration of the fundamentals of RFR, GBR and XGB, they showed experimentally that ensemble methods improved on support vector regression (SVR) for individual wind farm energy prediction.

GBRT based electricity price prediction model has been developed by Agrawal et al. **[2019]**. They also compared the model with relevance vector machine, multilayer perceptron and forest regression models. However, the developed model outperformed all of these models. GBRT has also been used for stock prediction **[Kohli et al., 2019]**. However, to the best of our knowledge, Gradient Boosted Regression Tree has not been applied to develop construction project duration and cost prediction model.

2.4. Research Gap

Although previous studies, to some extent, have developed many models for construction duration and cost prediction using different machine learning algorithms, ensemble type machine learning algorithms have not been applied yet in this area. Therefore, this study aims to use one of the recent ensemble machine learning algorithm Gradient Boosted Regression Tree (GBRT) for developing construction schedule and cost prediction models. Selection of significant features is an important step before building the models. However, in case of gradient boosted tree, optimization of hyper-parameter value is required to reduce the over fitting problem. There is a little work on the regularization of hyper-parameters of gradient boosted regression. Considering the research gaps mentioned above, this research aims to optimize the hyper-parameters of Gradient Boosted Regression Tree and to develop construction duration and cost prediction models based on regularized gradient boosted regression tree. Feature selection will be performed using one-way ANOVA F-test and the regularization of hyper-parameters will be done by Random Search algorithm.

CHAPTER 3: THEORETICAL BACKGROUND

In this study, theoretical background covers the topics of one-way ANOVA F-test, Gradient Boosted Regression Tree (GBRT) and regularization of hyper-parameters. This chapter explains the computational method of one-way ANOVA F-test and the Gradient Boosted Regression Tree (GBRT) algorithm. This chapter also discusses different hyper-parameters of GBRT and one of the methods of regularization, random search for selecting the best values of hyper-parameters.

3.1. One-way ANOVA F-test

Analysis of variance (ANOVA) is a statistical analysis tool that splits an observed aggregate variability found inside a dataset into two parts. One part is systematic factor and the other is random factor. The systematic factors have a statistical influence on the given dataset, while random factors do not. ANOVA test is used to determine the influence that independent variables have on the dependent variable in a regression study. One-way ANOVA uses F-test to assess whether the expected values of a quantitative variable within several pre-defined groups differ from each other. The formula for the one-way ANOVA F-test statistic is

$$F = \frac{explained\ variance}{unexplained\ variance} \tag{3.1}$$

Or,

$$F = \frac{between - group\ variability}{within - group\ variance} \tag{3.2}$$

The explained variance or between-group variability is

$$\sum_{i=1}^{K} n_i\,(\bar{Y}_i - \bar{Y})^2/(K-1) \tag{3.3}$$

Where $\bar{Y}_i$ denotes the sample mean in the i-th group, n_i is the number of observations in the i-th group, $\bar{Y}$ denotes the overall mean of the data, and K denotes the number of groups.

The unexplained variance or within-group variability is

$$\sum_{i=1}^{K}\sum_{j=1}^{n_i} \left(\bar{Y}_{ij} - \bar{Y}_i\right)^2/(N-K) \tag{3.4}$$

Where, $\bar{Y}_{ij}$ is the j-th observation in the j-th out of K groups and N is the overall sample size. This F-statistics follow the F-distribution with degrees of freedom $d_1 = K - 1$ and $d_1 = N - K$. The statistic will be large if the between group variability is large relative to the within-group variability, which is unlikely to happen if the population means of the groups all have the same value. In this study, one-way ANOVA F-test has been used for selecting the significant features. One-way ANOVA F-test method analyzes the data such that one response variable is calculated under various conditions identified by one input variable. It is often used in the analysis of data and drawing interesting information based on p-value. One-way ANOVA analysis by comparing the given dataset and returns a single p-value, which is significant. If the p-value is less than certain predefined value, then the input variable is considered to be significant. On the other hand, if p-value is greater than the predefined value, then the input variable is considered to be not significant.

3.2. Ensemble Machine Learning Algorithms

The ensemble algorithms consist of different basic models such as decision tree, neural networks, etc., and each basic models offers an alternative solution to the problem. The predictions of each model are somewhat combined to generate the final output model which is usually averaged or weighted. The combination of each model group predictions often results in a more stable and exact prediction than the prediction provided by each of the basic models in the group. Our everyday life often uses the basic idea behind the ensemble methods. In decision-making, it is usually used to seek opinions of others. By weighted combinations of these ideas, it can make more informed decisions. The effectiveness of ensemble methods depends largely on the variety of simple models that make up the ensemble. Combining results of several basic models is only useful if individual models have various outputs, or, in other words, disagree on some inputs with each other. Total error reduced by assembling methods by fixing error in each model. No advantage is that models which make similar mistakes are combined. The total error of the model can be reduced by combining individual base models that make various errors (or errors). Various models can be achieved through the application of different training data sets or through different training parameters for each model. Two popular ensemble techniques that use different sampling methods to create various training data for the acquisition of different basic models are packaging and boosting. Despite similar training datasets, the basic models are often compelled to be weak in order to produce various basic models. This can lead to various model results by disturbing the training data. Trees are one type of base model used for assemblies. They can be very prone to small changes in training data and a slight shift in training data will result in very different reversal trees. This unique property makes them successful assembly candidates. Moreover, trees are fast and easy algorithms which reduce the time and complexity of computation. Tree-based ensemble methods create many different trees and then combine the results of each tree. The benefit of an ensemble tree is that the variance can be minimized by means of average. Details will be illustrated in the later section. In general, the random forest and the gradient boosting regressive tree exist in two effective tree-based ensemble algorithms. Both approaches are based on a single regression tree. The random forest approach is derived from the principle of bagging, and boosting technique is the theoretical basis for gradient regression. A single regression tree is explained shortly in the following section, and then how various ensemble trees can be created **[Zhang & Haghani, 2015]**.

3.2.1 Single regression tree

A single tree model partitions the feature space into a set of regions and fits a simple model (a constant) for each region. For simplicity, consider a regression problem with continuous response variable Y and two independent variables X_1 and X_2. The space is first split into two regions and model the response Y (mean of Y) individually in each region. Then, it continues to split each individual region into two more regions and continue the process until some stopping rule is met. The Figure 3.1(a) shows that the feature space is divided into five regions $\{R_1, R_2, R_3, R_4, R_5\}$ according to two variables X_1 and X_2 using four split-points b_1, b_2, b_3 and b_4. The size of the tree is the total number of end nodes of the tree. In Figure 3.1(b), the size of the tree is 5 as the tree is partitioned into 5 regions or end nodes. During each partition process, the best fit is achieved through the selection of variables and split-point. Figure 3.1(b) is a binary tree representation of the same model.

It is now considered a generalized version of the above example: a regression problem consisting of p inputs with one response variable. For example, there are n observations, each observation consists of ($y_i,\ x_{i1}$,x_{i2},… x_{ij},…. x_{ip}) for $i = 1,2, \dots., n, j = 1,2, \dots, p$. In terms of construction schedule and cost prediction, y_i can be construction duration or cost. x_{i1}, x_{i2},…, x_{ij},…. x_{ip} are variables that are relevant to predicted construction schedule and construction cost, such as location, land size or other external factors. A single regression tree is the basic model for Random Forest and Gradient Boosting Regression Tree methods.

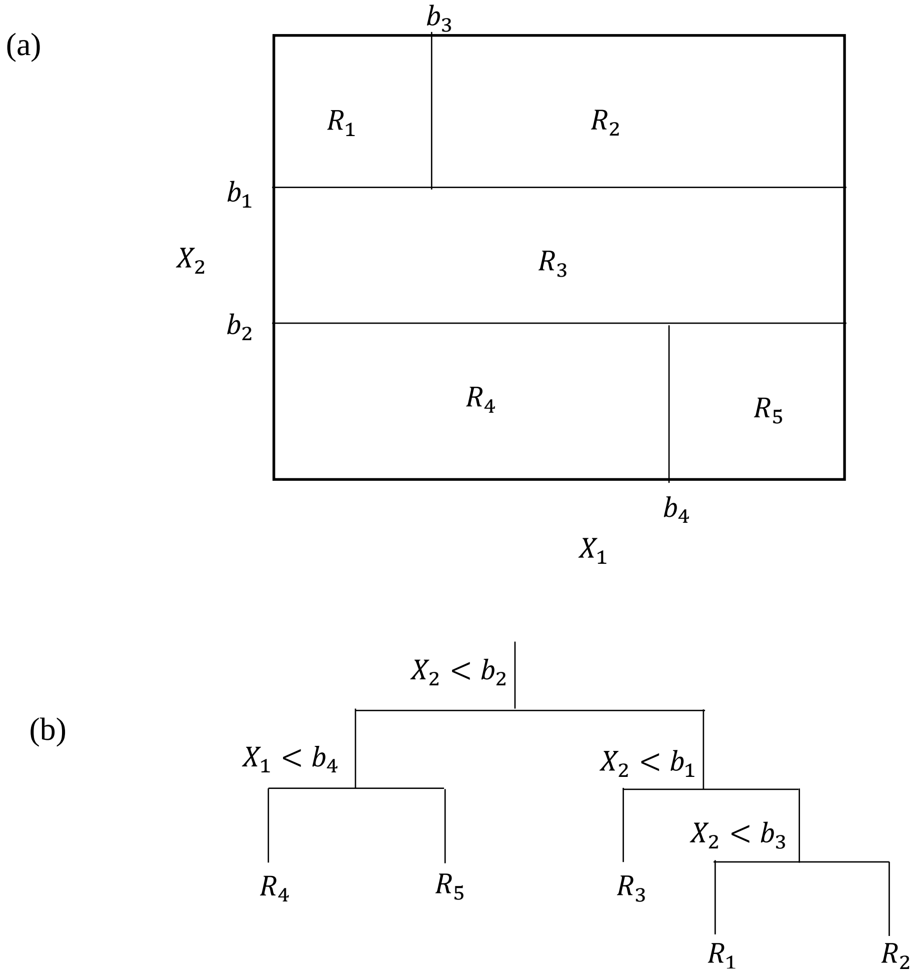


Figure 3.1: Single regression tree

3.2.2. Random Forest Algorithm

In 2001, Breiman **[2001]** developed Random Forest. It brings together two powerful machine learning techniques: Amit & Geman **[1997]** and Ho's **[1998]** idea of 'bagging' and their random features. The following sections clarify the idea of bagging and random selection of features to understand how a Random Forest works **[Zhang & Haghani, 2015]**.

In bagging or bootstrap aggregation, each individual based model is trained on the bootstrap sample from the training data. The bootstrap techniques were initially developed to estimate the sampling distribution by sampling the original data of an estimator from limited data. Bootstrap was used to create a variety of datasets for the training base model in recent development of ensemble techniques. For a given training data set with sample size n, bagging generates k new training set, each with sample size n, by sampling from the original training data set uniformly and with replacement. Through sampling with replacement, some observations appear more than once in the bootstrap sample, while other observations will be ‘left out’ of the sample. Then, k base models are trained using the newly generated k training set and combined through averaging (regression problem) or majority voting (classification problem).

The individual base model should have the characteristics of instability to ensure the effectiveness of the bagging process. Bagging increases prediction precision by the diversity of the basic models generated by a disturbed training set. The basic model should be poor to obtain different basic models from similar training sets. "Weak" here means a model that is somewhat better than random conjecture. The simplest and easiest ensemble tree, the bagged tree, can be accomplished with a tree as the basic model. Growing tree in the ensemble is generated by randomly drawing data samples that replace the original data. However, with lots of data, it is usually used to learn the same regression tree. Averaging output of these trees does not improve prediction accuracy.

Random selection of features is the second technique used by the random forest. Further development of the bagged regression tree is the random forest. The bootstrapped samples are still focused on the output of individual trees. It makes only a random subset of characteristics at each tree splitting node instead of using all features. Thus, variety between basic models is enhanced. The pseudo-code for random forest is shown in Table 3.1.

Random Forest increases predictive accuracy by combining several noisy yet almost neutral trees, by reducing variance. According to Hastie et al. **[2009]**, the variance of a Random Forest with total number (M) of trees is:

$$\rho\sigma^2 + \frac{1-\rho}{M}\sigma^2 \tag{3.5}$$

Where the variance of individual tree is indicated by σ^2, ρ denotes correlation between the trees, and K is the total number of trees in the ensemble. It is obvious that the second term tends to be zero by increasing the total number of trees M. Therefore, the variance of a Random Forest depends on three things:

(1) The correlation ρ between any pair of trees: the total variance is decreased by decreasing the correlation. This can be achieved by: randomly selecting ν out of the p variables to split at each splitting node when growing a tree on a bootstrapped dataset. Reducing ν, reduces both the correlation between trees and the strength of individual tree, and vice versa. Therefore, there is a need to find the optimal value of ν for certain dataset.

(2) The variance σ^2 of each individual tree, or in other words, the strength of each individual tree: Strengthening the performance of each individual tree can decrease the total variance of the model.

Table 3.1: Pseudo-code for Random Forest Ensemble algorithm **[Breiman, 2001]**

```
Algorithm-1: Random Forest Ensemble
Initialize the total number of trees $(M)$ to be generated and the number $\nu < \rho$ of variables used for each individual tree:
for $m = 1$ to $M$ do
    Draw a random sample $S^*$ of size $n$ with replacement from the original training data (This is also referred to as the bootstrap sample. This sample will be the training data to grow the tree);
    Grow a tree $T_m$ using training sample $S^*$ through the following loop:
        Do until (the maximum node size $nod_{min}$ is reached)
        for the terminal node of the tree;
            Randomly select $\nu$ variables out of the $p$ variables;
            Select the best pair of split variable/point among the $\nu$ variables;
            Split the node into two daughter nodes;
        end for
    Output the constructed tree $T_\nu(x)$;
end for
```

(3) The total number of trees M: The second term of the equation can be reduced by increasing M. Therefore, adequate number of trees should be trained to make sure the second term of the equation goes to zero.

Random Forests are usually based on the concept of bagging, but improve the random feature selection of each tree. The random forest theoretical history supports parallel computation, so parallel computing allows its training speed to be accelerated. Random forest predictions are mainly calculated by three variables, namely the association between each tree, its output and the total number of trees.

3.2.3 Gradient Boosted Regression Tree (GBRT)

Gradient Boosted Regression Tree has recently emerged as one of the top ensemble type machine learning algorithms. Trees are a base model form commonly used for ensemble machine learning techniques. Tree based algorithms are very fast and efficient as they reduce the computational time. Tree based ensemble techniques build a large number of different trees and then combine the results from each individual tree.

Unlike bagging, the boosting process sequentially produces simple models. The accuracy of prediction is improved by designing many sequence models by concentrating on those cases which are difficult to estimate. Examples difficult to estimate using previous simple models are more commonly found in training data than those properly calculated in the boosting process. That new base model attempts to correct the errors of its preceding base models. From the answer **[Schapire, 1990]** to Kearns' question is the beginning of the boosting technique: Is a set of weak learner equivalent to a single strong learner? A weaker leaner is an algorithm that only performs marginally better than a random guess; a stronger simple model is an arbitrarily well correlated prediction or classification algorithm. It is really important to address this issue. When contrast to a strong model, it is always easier to predict a weak model. Schapire **[1990]** proves that the response is good by using boosting algorithms to combine many weak models into one precise model. Table 3.2 represents the pseudo-code of boosting algorithm.

Table 3.2: Pseudo-code of Boosting algorithm **[Schapire, 1990]**.

```
Algorithm-2: Boosting Algorithm
Determine the total number of base models as $M$:
Define the initial training sample distribution as $D_1 = D$
for $m = 1$ to $M$ do
    Train a base model $B_m(x)$ from the training sample distribution $D_m$.
    Compute the error of the model.
    Adjust the distribution $D_m$ to $D_{m+1}$ to make the mistake of the model more evident.
    Output the constructed base model $B_m(x)$.
end for
Output the prediction of the ensemble trees for a given new input $x$: $\frac{1}{M}\sum_{m=1}^{M} B_m(x)$;
```

The key difference between the bagging and boosting approaches is that the boosting approach carefully tests the training data for each successive model to provide the most valuable details. The corrected distribution is based on the error produced by previous models during each training phase. The probability of selecting a single example is not equal to the boosting algorithm. Compared to the bagging approach where each sample is uniformly selected to produce a training dataset, the probability of selecting an individual sample is not equal for the boosting algorithm. Samples misclassified or misestimated are more likely with higher weight to be picked. Therefore, the samples that were misclassified by earlier models are highlighted in each new model.

The boost suits additional models, such as a squared error or an absolute failure which minimized a certain loss function averaged over training data. The loss function calculates the value expected by the true value. A forward-looking modeling approach is one of the tentative solutions to this problem. The forward-looking approach incorporates new basic models sequentially, without modifying current parameters and model coefficients. The boost method is a type of "decent functional gradient" in relation to the regression problem.

Boosting fits additional models that minimize a certain loss function averaged over the training data, such as a squared error or an absolute error. The loss function measures the amount the predicted value deviates from the true value. One of the approximate solutions to this problem is by using a forward stage-wise modeling approach. The forward stage-wise approach sequentially adds new base models without changing parameters and coefficients of models that have already been added. In terms of regression problem, the boosting method is a form of ‘‘functional gradient decent’’. It is an optimization technique that minimizes a certain loss function by adding a base model at each step that best reduces the loss function. Table 3.3 represents the pseudo-code for Gradient Boosted Regression Tree algorithm.

Table 3.3: Pseudo-code of Gradient Boosted Regression Tree (GBRT) algorithm **[Friedman, 2002]**.

```
Algorithm-3: Gradient Boosted Regression Tree
Initialize $H_0(x)$ to be constant, $H_0(x) = argmin_\rho \sum_{i=1}^{N} L(y_i, \rho)$.
for $t = 1\ to\ T$ do
    for $i = 1$ to $N$ do
        Compute the negative gradient
            $\tilde{y}_{it} = -\left[\frac{\partial L(y_i, H(\boldsymbol{x}_i))}{\partial H(\boldsymbol{x}_i)}\right]_{H(\boldsymbol{x})=H_{t-1(x)}}$
    end for
    Fit a regression tree to predict the targets $\tilde{y}_{it}$ from covariates $x_i$ for all training data.
    Update the model as
        $H_t(\boldsymbol{x}) = H_{t-1}(\boldsymbol{x}) + \nu \sum_{j=1}^{J_t} \gamma_{jt} 1(\boldsymbol{x} \in R_{jt})$
end for
Output the final model    $f_M(x)$
```

The computational steps for the generic gradient boosting method is as follows **[Friedman, 2002; Hastie et al., 2009]**:

Input features and target features are defined as $\boldsymbol{x} = (x_1, x_2, \dots \dots, x_k)$ and$y$, respectively. Let $\{y_i, \boldsymbol{x_i}\}_1^M$ be a set of training data including M pairs. The GBRT algorithm iteratively constructs T different regression trees $f(\boldsymbol{x}, \boldsymbol{a}_1), \dots \dots, f(\boldsymbol{x}, \boldsymbol{a_t})$from the set of training data and constructs the additive function $H(\boldsymbol{x})$ as follows:

$$H(\boldsymbol{x}) = \rho_0 + \sum_{t=1}^{T} \rho_t f(\boldsymbol{x}, \boldsymbol{a_t}) \tag{3.6}$$

Where, ρ_t and $\boldsymbol{a}_t$ are a weight and vector of parameters for the t^{th} regression tree $f(\boldsymbol{x}, \boldsymbol{a_t})$and $\rho_0$is an initial value. Both the weight ρ_t and the parameters $\boldsymbol{a_t}$ are iteratively determined from $t = 1$ to $t = T$ so that a loss function $L\big(y, H(\boldsymbol{x})\big)$ is minimized. Now, an additive function is defined which is combined from the first regression tree to the $(t-1)^{\text{th}}$ regression tree as $H_{t-1}(\boldsymbol{x})$. The weight ρ_t and the parameter $\boldsymbol{a_t}$ for the i^{th} regression tree is determined as follows:

$$(\rho_t, \boldsymbol{a_t}) = \underbrace{argmin}_{\rho, \boldsymbol{a}} \sum_{i=1}^{N} L\big(y_i, H_{t-1}(\boldsymbol{x_i}) + \rho f(\boldsymbol{x}_i, \boldsymbol{a})\big) \tag{3.7}$$

Where, $H_0(\boldsymbol{x})$ is an initial value and given by $H_0(\boldsymbol{x}) = \rho_0 = argmin_\rho \sum_{i=1}^{N} L(y_i, \rho)$.

However, in general, it is not straightforward to solve Eq. (2). Therefore, gradient boosted regression tree separately and approximately $(\rho_t, \boldsymbol{a_t})$ with a simple two-step procedure (Friedman, 2002). To estimate the parameters $\boldsymbol{a_t}$ for the regression tree, the function defined by the regression tree approximates a gradient with respect to the current function $h_{t-1}(\boldsymbol{x})$ in the sense of least-square error as follows:

$$\boldsymbol{a}_t = \underbrace{argmin}_{a} \sum_{i=1}^{N} \big(\tilde{y}_{it} - f(\boldsymbol{x}_i, \boldsymbol{a})\big)^2 \tag{3.8}$$

Where $\tilde{y}_{it}$ is the gradient and is given by,

$$\tilde{y}_{it} = -\left[\frac{\partial L\big(y_i, H(\boldsymbol{x}_i)\big)}{\partial H(\boldsymbol{x}_i)}\right]_{H(x)=H_{t-1(x)}} \tag{3.9}$$

When the t^{th} regression tree using the $\boldsymbol{a}_t$ has $J_t$leaf nodes, the regression tree is given by

$$f\left(\boldsymbol{x}, \{R_{\boldsymbol{jt}}\}_{j=1}^{Jt}\right) = \sum_{j=1}^{Jt} \tilde{y}_{jt} 1\big(\boldsymbol{x} \in R_{jt}\big) \tag{3.10}$$

Where, R_{jt} is a disjoint region that the j^{th} leaf node of the t^{th} regression tree defines. The step size of the gradient descent can be forwardly estimated using a line search on the loss function. Then, the model updating rule becomes

$$H_t(\boldsymbol{x}) = H_{t-1}(\boldsymbol{x}) + \nu \sum_{j=1}^{J_t} \gamma_{jt} 1\left(\boldsymbol{x} \in R_{jt}\right) \tag{3.11}$$

Where, $\gamma_{jt} = \rho_t \tilde{y}_{jt}$ and $0 < \nu < 1$is a shrinkage parameter to improve the generalization capability.

3.4. Performance measure of GBRT

The most common approached have been utilized to determine the estimation accuracy in Gradient Boosted Regression Tree (GBRT) are:

- Mean absolute error (MAE)
- Mean absolute percentage error (MAPE)
- Mean squared error (MSE)

Mean absolute error is one of many ways to quantify the difference between an estimated and the actual value of the projects being estimated. According to Willmott & Matsuura **[2005]** the MAE is relatively simple. It involves summing the magnitudes or absolute values of the errors to obtain the 'total error' and then dividing the total error by number of exemplars in the data set, n, it can be defined by the following formula:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{3.12}$$

Where, y_i is the predicted output by the Gradient Boosted Regression Tree model and x_i is the actual or desired output.

The mean absolute percentage error (MAPE) is a quantity used to measure how close forecasts or predictions are to the eventual outcomes, according to Principe **[2010]**, the MAPE is defined by the following formula:

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - x_i}{x_i} \right| \times 100\% \tag{3.13}$$

Where, y_i is the predicted output by the Gradient Boosted Regression Tree model and x_i is the actual or desired output. Note that this value can easily be misleading. For example, say that the output data is in the range of 0 to 100. For one exemplar, the desired output is 0.1 and the predicted output is 0.2. Even though the two values are quite close, the percent error for this exemplar is 100 **[Principe, 2010]**.

The mean squared error (MSE) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of squares of the errors that is, the average squared difference between the estimated values and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss. It is perhaps the simplest and common metric for regression evaluation, but also probably the least useful. It is defined by the following equation:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2 \times 100\% \tag{3.14}$$

Where, y_i is the predicted output by the Gradient Boosted Regression Tree model and x_i is the actual or desired output. The higher this value, the worse the model is. It is never negative, since the individual prediction-wise errors are squared before summing them, but would be zero for a perfect model.

3.5 Hyper-parameter of Gradient Boosted Regression Tree (GBRT)

In machine learning, a hyper-parameter is a parameter whose value is set before the learning process begins. By contrast, the values of other parameters are derived via training. Gradient Boosted Regression Tree (GBRT) is an ensemble type of machine learning algorithm. The major hyper-parameters of GBRT are described below:

a) Number of trees: The total number of trees in the sequence or ensemble. The averaging of independently grown trees in bagging and random forest makes it difficult to overfit with too many trees. However, Gradient Boosted Regression Tree (GBRT) functions differently as each tree is grown in sequence to fix up the past tree's mistakes. For example, in regression, GBRT will chase residuals as long as user allows them to. Also depending on the of the other hyper-parameters, GBRT often require many trees but since they can easily over fit, the optimal number of trees must be found that minimize the loss function of interest with cross validation.

b) Minimum number of leaf: Minimum number of leaf also controls the complexity of each tree. Higher values of this hyper-parameter help prevent a model from learning relationships which might be highly specific to the particular sample selected for a tree (overfitting) but smaller values can help with imbalanced target classes in classification problems.

c) Tree depth: Tree depth controls the depth of the individual trees. Typical values ranges from a depth of 3-15 but it is not uncommon to see a tree depth of 1. Smaller depth trees such as decision stumps are computationally efficient (but require more trees); however, higher depth trees allow the algorithm to capture unique interactions but also increase the risk of over-fitting.

d) Learning rate: Learning rate determines the contribution of each tree on the final outcome and controls how quickly the algorithm proceeds down the gradient decent (learns). Values range typically 0.01 to 0.4. Smaller values make the model robust to the specific characteristics of each individual trees, thus allowing it to generalize well. Smaller values also make it easier to stop prior to over-fitting. This hyper-parameter is also called shrinkage.

3.6 Random Search for Hyper-Parameter Regularization

Hyper-parameters have great impact on the performance of machine learning algorithms. Hyper-parameters have to be set before training the model. In Gradient Boosted Regression Tree, the values of some hyper-parameters such as number of trees, maximum depth, minimum sample leaf and learning rate have significant impact on model performance. Therefore, these hyper-parameters need to be optimized. Two popular hyper-parameter tuning algorithms are grid search and random search algorithm. Random Search algorithm has shown better performance than grid search algorithm in many experiments **[Bergstra & Bengio, 2012]**. In this study, Random Search algorithm has been used for hyper-parameter selection. In Random Search algorithm, the trials are given by randomly chosen the values of hyper-parameters within certain range. In this study, Random Search algorithm has been applied by using RandomSearchCV library in Pyhton 3.6.

The generic random search algorithm is described by a sequence of iterates $\{X_k\}$ on iteration $k = 0,1, \ldots$ which may depend on previous points and algorithmic parameters. The current iterate X_k may represent a single point, or a collection of points, to include population-based algorithms. The iterations are also capitalized to denote that they are random variables reflecting the probabilistic nature of the random search algorithm.

Step 0. Initialize algorithm parameters Θ_0, initial points $X_0 \subset S$ and iteration index $k = 0$.

Step 1. Generate a collection of candidate points $V_{k+1} \subset S$ according to a specific generator and associated sampling distribution.

Step 2. Update X_{k+1} based on the candidate points V_{k+1}, previous iterations and algorithmic parameters. Also update algorithmic parameters Θ_{k+1}.

Step 3. If a stopping criterion is met, stop. Otherwise increment K and return to Step 1.

CHAPTER 4: PROPOSED CONSTRUCTION SCHEDULE AND COST PREDICTION FRAMEWORK

This study proposes a framework for predicting the construction schedule and cost more accurately at the early stage for low rise, medium rise and high rise building. This proposed framework consists of several key phases which are: preliminary phase, data collection and preprocessing phase, feature selection phase, regularization and model development phase and result comparison phase. These key phases have to be followed to predict the construction schedule and cost more accurately as shown in Figure 4.1. The details of the proposed framework are described below:

4.1 Preliminary Phase

The first step of developing construction schedule and cost prediction is to identify the input features required to develop the models. After reviewing previous literature and consulting with experts, 10 input features were selected for building the schedule and cost prediction models. These 10 features are: “Location (F1)”, “Land size (F2)”, “Floor height (F3)”, “Floor area (F4)”, “No. of basement (F5)”, “Design (F6)”, “Finishing materials type (F7)”, “Approval complexity (F8)”, “Workforce level (F9)” and “No. of floor (F10)” Table 4.1 represents the description of these 10 features.

4.2 Data Collection and Data Preprocessing Phase

After identifying the input features for the prediction model, historical data of those feature will be collected in this step. Data collection is a crucial step in machine learning model development. Data should be collected from the same distribution so that the learning process works well. In the data preprocessing step, statistical analysis will be performed on the collected data. Data encoding process will be done before model development. This step is required to understand the nature of the collected data.

Table 4.1: Description of input features for schedule and cost prediction models

No.	Feature name	Brief description	Source
1	Location (F1)	Location refers to the place of construction of building and infrastructure	**[Koo et al., 2010; Bala et al., 2014]**
2	Land size (F2)	Land size means the area of the land where the construction work is being commenced.	Experts
3	Floor height (F3)	Floor height refers to height measured from the top of floor to the surface of the ceiling plus the thickness of the floor between the planes.	**[Abu Hammad et al., 2008]**
4	Floor area (F4)	In construction, floor area refers to the area of each floor measured to the external face of the external walls.	**[Sonmez, 2011; Wang et al., 2012]**
5	No. of basement (F5)	No. of basement refers to the no. of floor of a building which is partly or entirely below ground level.	Experts
6	Design (F6)	Design refers to the architectural design of the infrastructure.	**[Lowe et al., 2006; Jin et al., 2012]**
7	Finishing materials type (F7)	Finishing materials type is the type of those materials and items used to improve the service and decorative qualities of buildings and structures.	**[Wang et al., 2012; Kim et al., 2013]**
8	Approval complexity (F8)	Approval complexity indicates the complication regarding the approval of building designs and other documents from different government agencies.	Experts
9.	Workforce level (F9)	Workforce refers to the required level of manpower for completing the construction project.	Experts
10.	No. of floor (F10)	No. of floor refers to the storey of the building from the ground floor to top floor.	**[Sonmez, 2011]**

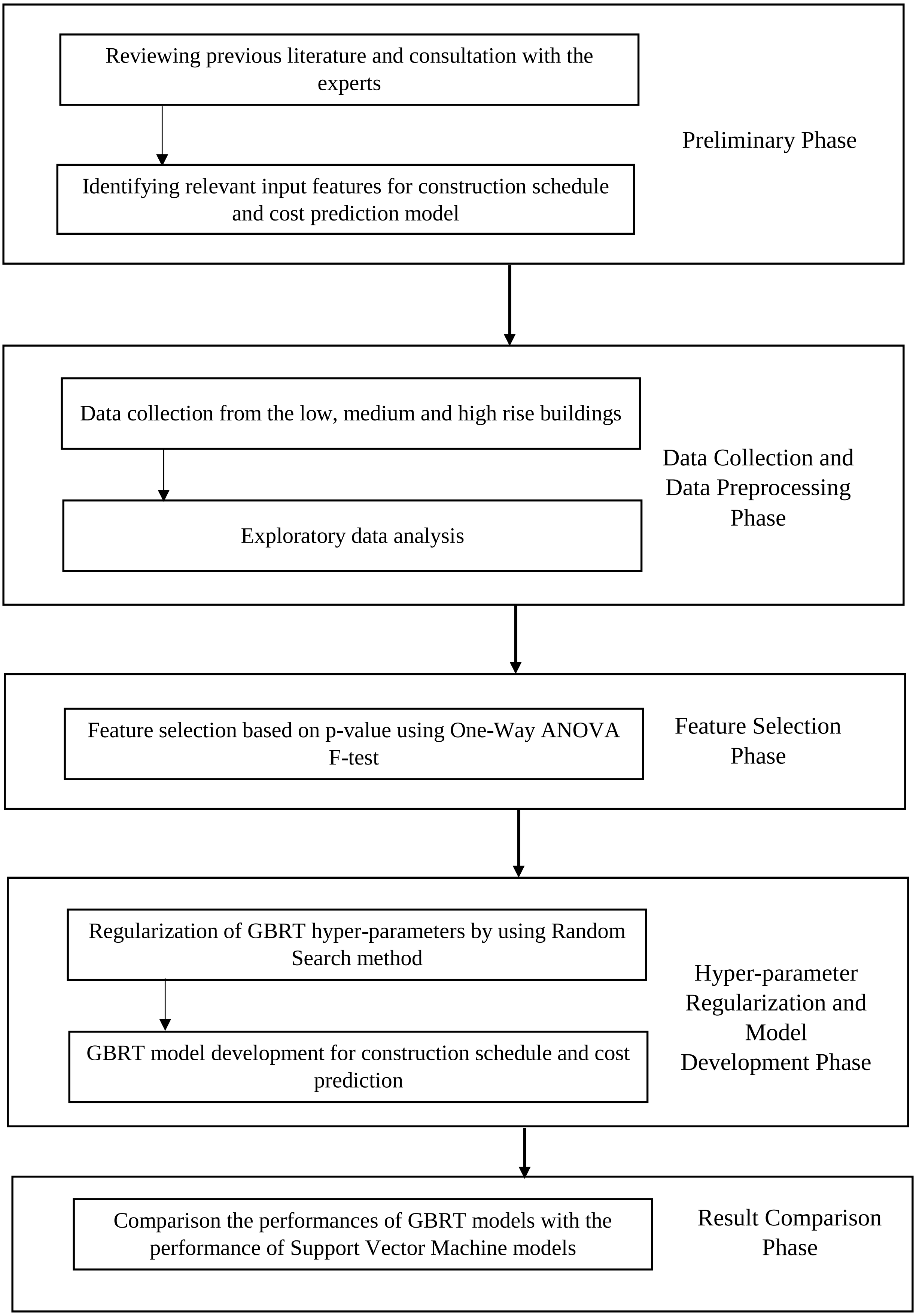

Reviewing previous literature and consultation with the experts
Identifying relevant input features for construction schedule and cost prediction model
Preliminary Phase
Data collection from the low, medium and high rise buildings
Exploratory data analysis
Data Collection and Data Preprocessing Phase
Feature selection based on p-value using One-Way ANOVA F-test
Feature Selection Phase
Regularization of GBRT hyper-parameters by using Random Search method
GBRT model development for construction schedule and cost prediction
Hyper-parameter Regularization and Model Development Phase
Comparison the performances of GBRT models with the performance of Support Vector Machine models
Result Comparison Phase

Figure 4.1: Flow chart of the current research

4.3 Feature Selection Phase

In this phase, statistically significant features have been identified through feature selection process. In this study, one-way ANOVA F-test has been used as feature selection method. One-way ANOVA F-test has been performed to select those features which are statistically significant based on p-value. Features with p-values less than 0.05 should be selected. In this study, p-value was set to 0.05. Any value lesser than 0.05 was considered as effective feature while any value greater than this value was considered as non-significant.

4.4 Hyper-parameter Regularization and Model Development Phase

4.4.1 Hyper-parameter Regularization

Hyper-parameters can be classified as model hyper-parameters, that cannot be inferred while fitting the machine to the training set because they refer to the model selection task, or algorithm hyper-parameters. To develop the Gradient Boosted Regression Tree (GBRT) model, the best values of hyper-parameters have been derived in this step. This process is known as hyper-parameter regularization. Four hyper-parameters of GBRT have been regularized. These four hyper-parameters are: number of trees, minimum sample leaf, tree depth and learning rate. To regularize these four hyper-parameters, Random Search method has been used. In this study, regularization has been performed using RandomSeachCV library in Python 3.6.

4.4.2 Model Development

After finding the best values of hyper-parameters, the construction schedule and cost prediction models have been developed using GBRT. The dataset was split into training set and testing set. The model was trained with the training data and was tested with the testing data. The mathematical expression of GBRT model.

Inputs for the model:

The two inputs for GBRT model are: (a) Dataset $\{x_i, y_i\}_{i=1}^{n}$ where x_i denotes the i^{th} row and y_i denotes the actual output of the i^{th} row and n is the number of rows in the dataset, and (b) Loss function $L\big(y_i, F(x)\big)$ where $F(x)$ is the predicted value. In this study, L_2 loss function or Least Square Error loss function is used.

Building the model:

Step 1: Initialize the model with a constant value: $F_0(x) = \underbrace{argmin}_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma)$; here, γ is the predicted value for the model.

Step 2: For $m = 1$ to M, compute $r_{im} = -\left[\frac{dL(y_i, F(x_i))}{dF(x_i)}\right]_{F(x)=F_{m-1}(x)}$, Here, M is the number of trees in the model. This is one of the hyper-parameters of GBRT. The value of number of trees in the model (*M*) has been selected using Random Search method. r_{im} is the pseudo residual for i^{th} row and *m* tree.

Step 3: For $m = 1$ to M, fit a regression tree to the r_{im} values and create terminal regions R_{jm} for $j = 1,2, \dots, j_m$. Here, j is the leaf number.

Step 4: For $j = 1,2, \dots j_m$, compute $\gamma_{jm} = \underbrace{argmin}_{\gamma} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$.

Step 5: Updating rule: $F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{j_m} \gamma_{jm} I(x \in R_{jm})$. Here, ν is the learning rate. To compare the results from Gradient Boosted Regression Tree models for construction schedule and cost prediction, two performance metrics namely mean squared error (MSE) and mean absolute percentage error (MAPE) have been used. The performance of GBRT models were compared with the results from Support Vector Machine and Multiple Linear Regression models.

CHAPTER 5: NUMERICAL EXPERIMENTATION

5.1 Data Collection

As machine learning models are developed based on some datasets, data collection is the primary stage of developing a machine learning model. This step is very crucial because most of the cases, the quantity and quality of data determine the accuracy of the corresponding machine learning model. In this study, the aim is to develop construction schedule and cost prediction model based on a popular ensemble machine learning algorithm named Gradient Boosted Regression Tree (GBRT). However, to get better performance, the hyper-parameters of the models have been regularized.

To develop the model, data for these features were collected from a reputed real estate development company. The data consists of historical information about 69 residential construction projects from a reputed construction engineering company of Dhaka city of Bangladesh executed from 2013 to 2018. These 69 residential projects were divided into low rise, medium rise and high rise building. Table 5.1 represents the data description for the schedule and cost prediction models. The illustrative example has the following conditions and assumptions:

i. All the data were collected from the same distribution. Here, the same distribution means the same company. There is some true but unknown data distribution from which each of the training and test points are drawn independently.
ii. All the recent projects were considered for building schedule and cost prediction models.
iii. All the projects were categorized as low rise, medium rise and high rise building. Buildings having the number of floors between 8 to 10, were considered as low rise buildings. Medium rise buildings have number of floors between 11 to 13. Buildings with number of floors between 14 to 16 were considered as high rise buildings.
iv. External non-controllable factors such as political turbulence, environmental factors, country’s economic conditions were not considered for developing the models.
v. Some of the features have very low effect on the response variables. They will be removed by using feature selection methods.
vi. The raw data were converted to numerical form using data encoding method.

Table 5.1: Description of data for schedule and cost prediction models.

Feature type	Feature name	Description
Input	Location (F1)	1 = Dhanmondi, 2 = Gulshan, 3 = Banani, 4 = Lalmatia, 5= Baridhara
Input	Land size (F2)	1 = 335 – 670 m^2, 2 = 670 – 1005 m^2, 3 = 100 – 1340 m^2
Input	Floor height (F3)	1 = 3.1 m, 2 = 3.4 m, 3 = 3.7 m
Input	Floor area (F4)	1 = 230 – 334 m^2, 2 = 334 – 390 m^2, 3 = 390 – 450 m^2
Input	No. of basement (F5)	1 = 1 basement, 2 = 2 basements, 3 = 3 basements
Input	Design (F6)	1 = Design type C, 2 = Design type B, 3 = Design type C
Input	Finishing materials type (F7)	1 = Finishing materials type C, 2 = Finishing materials type B, 3 = Finishing materials type A
Input	Approval complexity (F8)	1 = Low, 2 = Medium, 3 = High
Input	Workforce level (F9)	1 = Very low level, 2 = Low level, 3 = Medium level, 4 = High level, 5 = Very high level.
Input	No. of floor (F10)	8 to 16 (Integer)
Output	Duration (F11)	26 to 60 months (Continuous)
Output	Cost (F12)	BDT 12.5 to 23.5 (Continuous)

5.2 Exploratory Data Analysis

Among these features, “Location (F1)”, “Land size (F2)”, “Floor height (F3)”, “Floor area (F4)”, “No. of basement (F5)”, “Design (F6)”, “Finishing material type (F7)”, “Approval complexity (F8)”, “Workforce level (F9)” and “No. of floor (F10)” were considered as input features. In the feature selection method, one-way ANOVA-F test has been used to choose the most significant predictor features for developing the model. In this study, two separate models have been built to predict construction time duration and cost for low rise, medium rise and high rise buildings. Two features namely “Duration (F11)” and “Cost (F12)” were used as output or target features for these models respectively.

Table 5.2 represents the quantity of different types of features for low rise buildings. There are 20 buildings in the low rise category. Form the table, it is seen that there are 5 types of “Location (F1)”. “Floor height (F3)”, “Floor area (F4)”, “No. of basement (F5)”, “Design (F6)”, “Finishing materials type (F7)”, and “Approval complexity (F8)” have been divided into 3 types. Again, there are 2 types of “Land size (F2)” and 4 types of “Workforce level (F9)” for low rise buildings.

Table 5.2: Types of features for low rise buildings.

	Type 1	Type 2	Type 3	Type 4	Type 5
Location (F1)	5	3	8	2	2
Land size (F2)	17	3	0	0	0
Floor height (F3)	8	8	4	0	0
Floor area (F4)	13	4	3	0	0
No. of basement (F5)	7	11	2	0	0
Design (F6)	13	6	1	0	0
Finishing materials (F7)	9	5	6	0	0
Approval complexity (F8)	7	5	8	0	0
Workforce level (F9)	5	3	8	4	0

Table 5.3 represents the quantity of different types of features for medium rise buildings. There are 28 buildings in the medium rise category. Form the table, it is seen that there are 5 types of “Location (F1)” in case of medium rise buildings. However, “Land size (F2)” and “No. of basements (F5)” are divided into 2 types. Again, “Floor height (F3)”, “Floor area (F4)”, “Design (F6)”, “Finishing materials type (F7)”, and “Approval complexity (F8)” have been divided into 3 types. There are 4 types of “Workforce level (F9)” for medium rise buildings.

Table 5.3: Types of features for medium rise buildings.

	Type 1	Type 2	Type 3	Type 4	Type 5
Location (F1)	10	6	3	6	3
Land size (F2)	0	19	9	0	0
Floor height (F3)	8	8	12	0	0
Floor area (F4)	14	11	3	0	0
No. of basement (F5)	0	12	16	0	0
Design (F6)	12	9	7	0	0
Finishing materials (F7)	11	8	9	0	0
Approval complexity (F8)	7	11	10	0	0
Workforce level (F9)	0	3	7	11	7

The quantity of different types of features for high rise buildings have been shown in Table 5.4. There are 21 buildings in the high rise category. Here, also the “Location (F1)” feature is divided into 5 types. Like medium rise buildings, “Land size (F2)”, “Floor area (F4)” and “No. of basements (F5)” are divided into 2 types for high rise buildings. “Floor height (F3)”, “Design (F6)”, “Finishing materials type (F7)”, and “Approval complexity (F8)” have been divided into 3 types. “Workforce level (F9)” is of 4 types for high rise buildings.

Table 5.4: Types of features for high rise buildings.

	Type 1	Type 2	Type 3	Type 4	Type 5
Location (F1)	3	7	1	2	8
Land size (F2)	0	3	18	0	0
Floor height (F3)	7	4	10	0	0
Floor area (F4)	0	8	13	0	0
No. of basement (F5)	0	10	11	0	0
Design (F6)	10	4	7	0	0
Finishing materials (F7)	7	7	7	0	0
Approval complexity (F8)	8	7	6	0	0
Workforce level (F9)	0	0	2	6	13

5.3 Kendall’s Tau Correlation Coefficient

The Kendall’s Tau correlation coefficient, commonly referred to as Kendall's τ coefficient (after the Greek letter τ, tau), is a statistic used to measure the ordinal association between two measured quantities. A τ test is a non-parametric hypothesis test for statistical dependence based on the τ coefficient. In this study, Kendall’s τ coefficient has been used to measure the correlation between the features. Table 5.5 shows the Kendall’s Tau correlation coefficient among the features. From the table, it is seen that there does not exist significant correlation between any two features. This means all the features are independent.

Table 5.5: Kendall’s Tau correlation coefficient for all the features

	F1	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**	**F10**	**F11**	**F12**
F1	1.000	0.194	0.425	0.023	0.086	-0.024	0.056	0.083	0.154	0.149	0.147	0.156
F2	0.194	1.000	0.184	0.461	0.491	0.134	0.055	-0.007	0.730	0.347	0.768	0.765
F3	0.425	0.184	1.000	0.267	0.293	0.021	0.184	0.053	0.152	0.126	0.129	0.162
F4	0.023	0.461	0.267	1.000	0.410	0.145	0.033	-0.017	0.463	0.378	0.413	0.427
F5	0.086	0.491	0.293	0.410	1.000	-0.104	0.010	-0.057	0.502	0.445	0.431	0.419
F6	-0.024	0.134	0.021	0.145	-0.104	1.000	0.174	0.046	0.041	0.053	0.111	0.103
F7	0.056	0.055	0.184	0.033	0.010	0.174	1.000	0.020	0.153	0.098	0.094	0.143
F8	0.083	-0.007	0.053	-0.017	-0.057	0.046	0.020	1.000	0.031	-0.024	0.013	-0.008
F9	0.154	0.730	0.152	0.463	0.502	0.041	0.153	0.031	1.000	0.472	0.697	0.760
F10	0.149	0.447	0.126	0.378	0.445	0.053	0.098	-0.024	0.672	1.000	0.812	0.733
F11	0.147	0.468	0.129	0.413	0.431	0.111	0.094	0.013	0.697	0.312	1.000	0.716
F12	0.156	0.565	0.162	0.427	0.419	0.103	0.143	-0.008	0.760	0.433	0.716	1.000

5.4 Feature Selection

In this stage of the study, feature selection has been performed on the collected data using one-way ANOVA for each category of building. Although, the primary stage (Data collection stage) has selected 10 input or predictive features for developing construction duration and cost prediction model, one-way ANOVA F-test has been performed to select those features which are statistically significant based on p-value. As suggested by Kumar et al. **[2015]**, features with p-values less than 0.05 should be selected. On the other hand, if the p-value is larger than 0.05 then the feature value was considered as non-significant. One-way ANOVA F-test was applied for both duration and cost prediction models separately for each category of building. Table 5.6 shows the result of one-way ANOVA F-test performed on the collected data of low rise building for dependent variable “Duration (F11)”. In case of duration prediction of low rise building, p-value for “Location (F1)” is 0.006 which is less than 0.05. The feature is statistically significant as the value is less than 0.05. Hence, this feature has been considered for developing duration prediction model for low rise building. Similarly, p-values for “Land size (F2)”, “Floor height (F3)”, “Floor area (F4)”, “No. of basement (F5)”, “Workforce level (F9)” and “No. of floor (F10)” are less than 0.05 which confirm that all these features are significant for the duration prediction of low rise building. So, all these features have been considered for model development. On the other hand, p-values for “Design (F6)”, “Finishing materials type (F7)”, and “Approval complexity (F8)” are above 0.05. This indicates that all these features are not statistically significant. Hence, all these features have been rejected.

Table 5.6: Results of one-way ANOVA performed on the collected data of low rise buildings for dependent variable “Duration (F11)”.

Feature	Sum of squares	Degree of freedom	Mean square	F	P-value	Status
Location (F1)						
Between Groups	145.133	4	36.283	5.516	0.006	Selected
Within Groups	98.667	15	6.578			
Total	243.800	19	_			
Land size (F2)						
Between Groups	176.020	1	176.020	46.745	0.000	Selected
Within Groups	67.78	18	3.766			
Total	243.800	19	_			
Floor Height (F3)						
Between Groups	140.300	2	70.150	11.522	0.001	Selected
Within Groups	103.500	17	6.088			
Total	243.800	19	_			
Floor area (F4)						
Between Groups	203.073	2	101.536	42.382	0.000	Selected
Within Groups	40.727	17	2.396			
Total	243.800	19	_			
No. of basement (F5)						
Between Groups	186.800	2	93.400	27.856	0.000	Selected
Within Groups	57.000	17	3.353			

Total	243.800	19	–			
Design (F6)						
Between Groups	24.108	2	12.054	0.933	0.413	Rejected
Within Groups	219.692	17	12.923			
Total	243.800	19	–			
Finishing materials type (F7)						
Between Groups	8.244	2	4.122	0.298	0.746	Rejected
Within Groups	235.556	17	13.856			
Total	243.800	19	–			
Approval complexity (F8)						
Between Groups	4.071	2	2.036	0.144	0.867	Rejected
Within Groups	239.729	17	14.102			
Total	243.800	19	–			
Workforce level (F9)						
Between Groups	197.800	2	98.900	36.550	0.000	Selected
Within Groups	46.000	17	2.706			
Total	243.800	19	–			
No. of floor (F10)						
Between Groups	135.600	2	67.800	10.652	0.001	Selected
Within Groups	108.200	17	6.365			
Total	243.800	19	–			

Table 5.7 represents the result of one-way ANOVA F-test performed on the collected data for dependent variable “Cost (F11)” of low rise buildings. In case of cost prediction of low rise buildings, the p-value of “Location (F1)” is 0.005 which is less than 0.05. Thus this feature has been selected. P-values for “Land size (F2)”, “Floor height (F3)”, “Floor area (F4)”, “No. of basement (F5)”, “Workforce level (F9)” and “No. of floor (F10)” are close to zero for cost prediction which indicate that these features are statistically significant. Thus, these features have been selected for developing construction cost prediction model of low rise building. On the other hand, p-values for “Design (F6)”, “Finishing material type (F7)”, and “Approval complexity (F8)” are more than 0.05. This confirms that these features are not statistically significant. For this reason, these features have been rejected.

Table 5.7: Results of one-way ANOVA performed on the collected data of low rise buildings for dependent variable “Cost (F12)”.

Feature	Sum of squares	Degree of freedom	Mean square	F-value	P-value	Status
Location (F1)						
Between Groups	32.643	4	8.161	5.758	0.005	Selected
Within Groups	21.259	15	1.417			
Total	53.902	19	_			
Land size (F2)						
Between Groups	41.794	1	41.794	62.134	0.000	Selected
Within Groups	12.108	18	0.673			
Total	53.902	19	_			
Floor Height (F3)						
Between Groups	35.042	2	17.521	15.793	0.000	Selected
Within Groups	18.860	17	1.109			
Total	53.902	68	_			
Floor area (F4)						
Between Groups	46.227	2	23.113	51.196	0.000	Selected
Within Groups	7.675	17	0.451			
Total	53.902	19	_			
No. of basements (F5)						
Between Groups	38.908	2	19.454	22.056	0.000	Selected
Within Groups	14.994	17	0.882			
Total	53.902	19	_			
Design (F6)						
Between Groups	5.944	2	2.972	1.054	0.370	Rejected
Within Groups	47.958	17	2.821			
Total	53.902	19	_			
Finishing materials type (F7)						
Between Groups	1.979	2	0.990	0.324	0.728	Rejected
Within Groups	51.923	17	3.054			
Total	53.902	19	_			

Approval complexity (F8)						
Between Groups	1.590	2	0.795	0.258	0.775	Rejected
Within Groups	52.312	17	3.077			
Total	53.902	68	_			
Workforce level (F9)						
Between Groups	45.299	2	22.649	44.755	0.000	Selected
Within Groups	8.603	17	0.506			
Total	53.902	19	_			
No. of floor (F10)						
Between Groups	42.386	2	21.193	31.285	0.000	Selected
Within Groups	11.516	17	0.677			
Total	53.902	19	_			

Figure 5.1 and 5.2 show the main effect plot of duration and cost prediction for low rise buildings, respectively. From the main effect plot, it is seen that the lines for “Design (F6)”, “Finishing materials type (D7)” and “Approval complexity (F8)” are almost horizontal indicating no effect on the response feature. On the other hand, the lines for the rest of the features are not horizontal indicating main effect on the response feature.

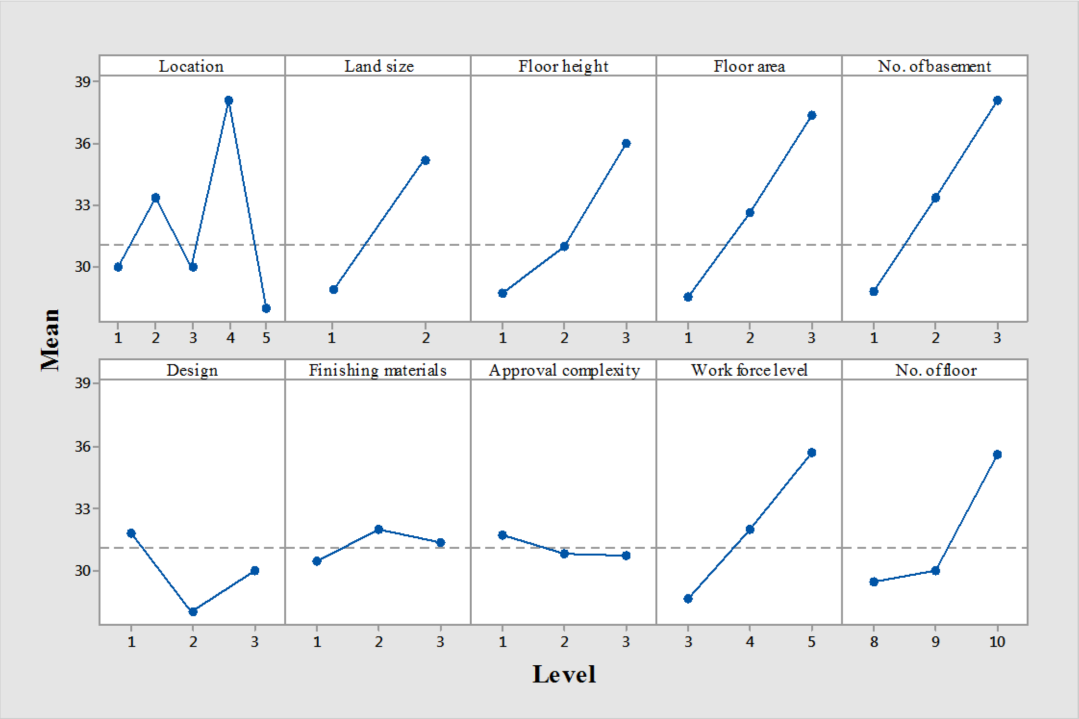


Figure 5.1: Main effect plot for “Duration (F11)” prediction of low rise buildings.

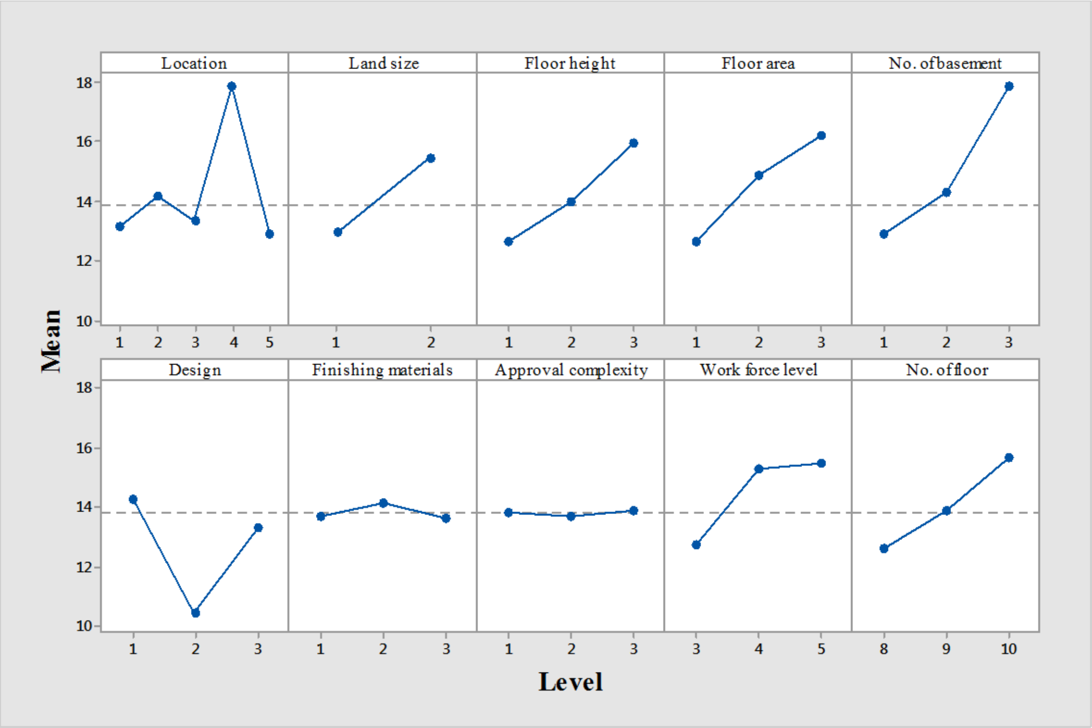


Figure 5.2: Main effect plot for "Cost (F12)" prediction of low rise buildings.

Table 5.8 shows the result of one-way ANOVA F-test performed on the collected data of medium rise buildings for dependent variable "Duration (F11)". In case of duration prediction of medium rise buildings, p-values for "Land size (F2)", "Floor area (F4)", "No. of basement (F5)", "Workforce level (F9)" and "No. of floor (F10)" are less than 0.05. Therefore, all of these features are significant for the duration prediction of medium rise buildings. These features have been selected for model development. On the other hand, p-values for "Location (F1)", "Floor height (F3)", "Design (F6)", "Finishing materials type (F7)", and "Approval complexity (F8)" are above 0.05. Therefore, all of these features were not considered for duration prediction of medium rise building.

Table 5.8: Results of one-way ANOVA performed on the collected data of medium rise buildings for dependent variable "Duration (F11)".

Feature	Sum of squares	Degree of freedom	Mean square	F	P-value	Status
Location (F1)						
Between groups	179.412	4	44.853	2.496	0.071	Rejected
Within Groups	413.267	23	17.968			
Total	592.679	27	_			

Land size (F2)						
Between Groups	142.152	1	142.152	8.204	0.008	Selected
Within Groups	450.526	26	17.328			
Total	592.679	27	_			
Floor Height (F3)						
Between Groups	16.529	2	8.214	0.356	0.704	Rejected
Within Groups	576.250	25	23.050			
Total	592.679	27	_			
Floor area (F4)						
Between Groups	231.513	1	231.513	16.666	0.000	Selected
Within Groups	361.166	26	13.891			
Total	592.679	27	_			
No. of basement (F5)						
Between Groups	278.679	1	278.679	23.075	0.000	Selected
Within Groups	314.000	26	12.077			
Total	592.679	27	_			
Design (F6)						
Between Groups	57.583	2	28.792	1.345	0.279	Rejected
Within Groups	535.095	25	21.404			
Total	592.679	27	_			
Finishing materials type (F7)						
Between Groups	56.258	2	28.129	1.311	0.287	Rejected
Within Groups	536.420	25	21.457			
Total	592.679	27	_			
Approval complexity (F8)						
Between Groups	9.882	2	4.941	0.212	0.810	Rejected
Within Groups	582.796	25	23.312			
Total	592.679	27	_			
Workforce level (F9)						
Between Groups	271.324	3	90.775	6.801	0.002	Selected
Within Groups	320.355	24	13.348			
Total	592.679	27	_			

No. of floor (F10)						
Between Groups	286.779	2	67.800	10.652	0.001	Selected
Within Groups	305.900	17	6.365			
Total	592.679	19	_			

Table 5.9 represents the result of one-way ANOVA F-test performed on the collected data for dependent variable “Cost (F11)” of medium rise buildings. In case of cost prediction of medium rise buildings, the p-values for “Land size (F2)”, “Floor area (F4)”, “No. of basement (F5)”, “Workforce level (F9)” and “No. of floor (F10)” are less than 0.05 which indicate that these features are statistically significant. On the other hand, p-values for “Location (F1)”, “Floor height (F3)”, “Design (F6)”, “Finishing material type (F7)”, and “Approval complexity (F8)” are larger than 0.05. This confirms that these features are not statistically significant.

Table 5.9: Results of one-way ANOVA performed on the collected data of medium rise buildings for dependent variable “Cost (F12)”.

Feature	**Sum of squares**	**Degree of freedom**	**Mean square**	**F**	**P-value**	**Status**
Location (F1)						
Between Groups	39.843	4	9.961	4.461	0.408	Rejected
Within Groups	51.358	23	2.233			
Total	91.201	27	_			
Land size (F2)						
Between Groups	34.139	1	34.139	15.555	0.001	Selected
Within Groups	57.062	26	2.195			
Total	91.201	27	_			
Floor Height (F3)						
Between Groups	7.516	2	3.758	1.123	0.341	Rejected
Within Groups	83.685	25	3.347			
Total	91.201	27	_			
Floor area (F4)						
Between Groups	42.240	1	42.240	22.431	0.000	Selected
Within Groups	48.961	26	1.883			
Total	91.201	27	_			

No. of basement (F5)						
Between Groups	29.052	1	29.052	12.154	0.002	Selected
Within Groups	62.149	26	2.390			
Total	91.201	27	–			
Design (F6)						
Between Groups	5.124	2	2.562	0.744	0.485	Rejected
Within Groups	86.077	25	3.443			
Total	91.201	27	–			
Finishing materials type (F7)						
Between Groups	11.535	2	5.767	1.810	0.184	Rejected
Within Groups	79.666	25	3.187			
Total	91.201	27	–			
Approval complexity (F8)						
Between Groups	2.481	2	1.240	0.350	0.708	Rejected
Within Groups	88.720	25	3.549			
Total	91.201	27	–			
Workforce level (F9)						
Between Groups	67.930	3	22.643	23.353	0.000	Selected
Within Groups	23.271	24	0.970			
Total	91.201	27	–			
No. of floor (F10)						
Between Groups	45.775	2	22.887	12.596	0.000	Selected
Within Groups	45.427	25	1.817			
Total	91.201	27	–			

Figure 5.3 and 5.4 represent the main effect plot of duration and cost prediction for medium rise building, respectively. From the main effect plot, it is seen that the lines for “Land size (F2)”, “Floor area (F4)”, “No. of basement (F5)”, “Workforce level (F9)”, and “No. of floor (F10)” are not horizontal which indicate that there exists main effect on the response feature. On the other hand, the lines for the rest of the features are almost horizontal indicating no main effect on the response feature.

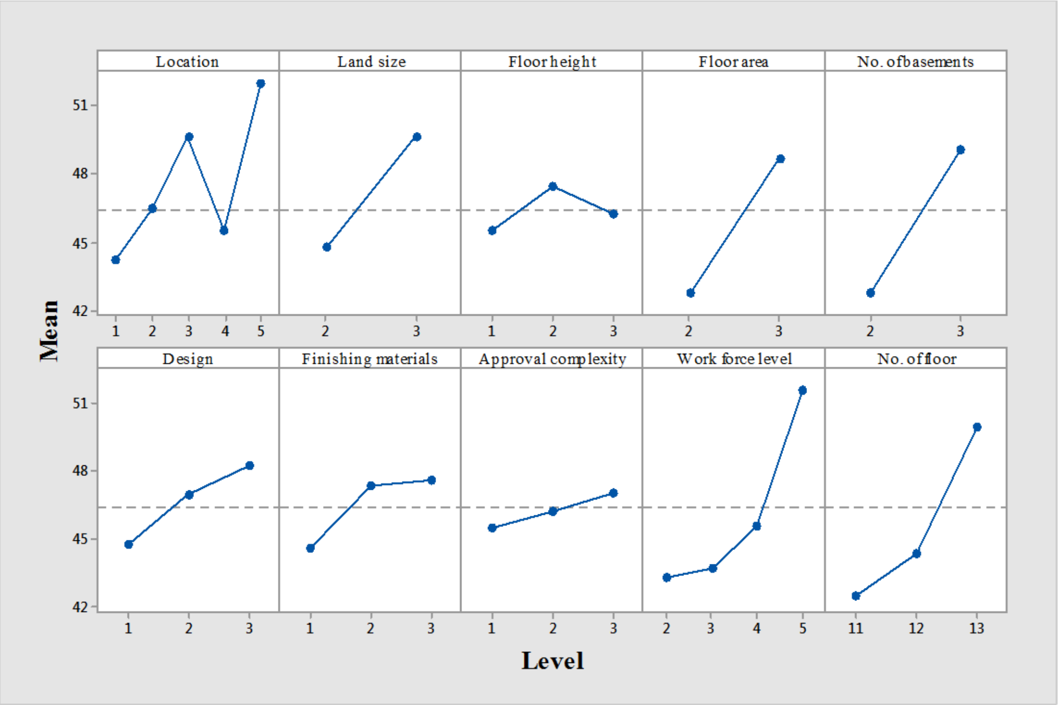


Figure 5.3: Main effect plot for “Duration (F11)” prediction of medium rise buildings.

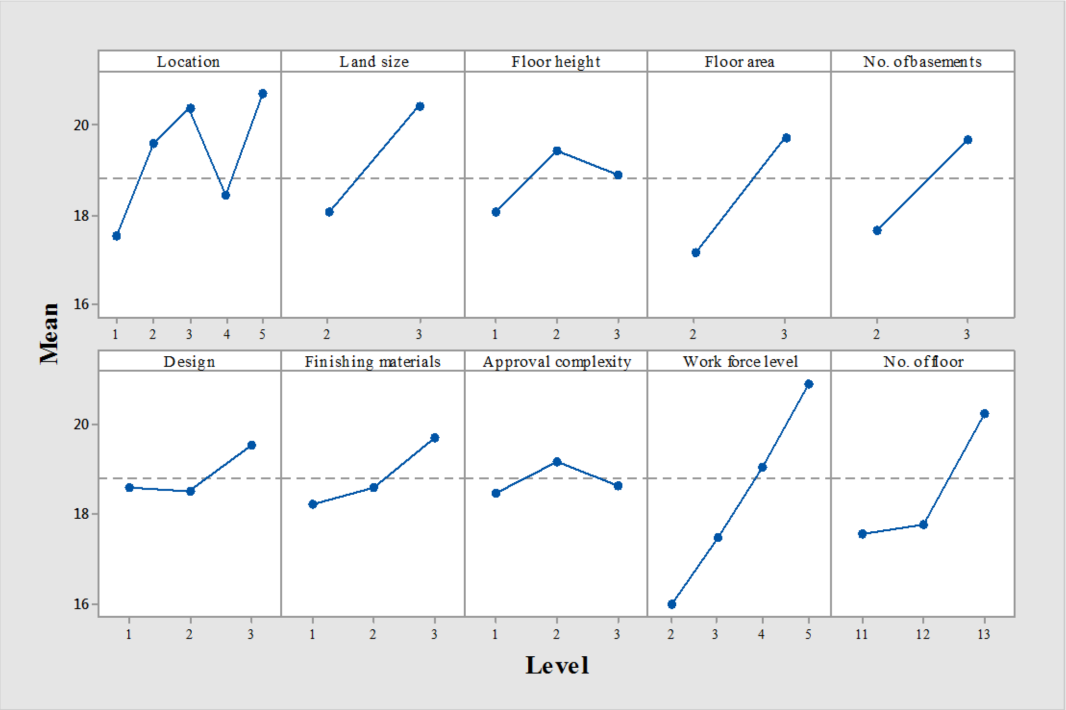


Figure 5.4: Main effect plot for “Cost (F12)” prediction of medium rise buildings.

Table 5.10 shows the result of one-way ANOVA F-test performed on the collected data of high rise buildings for dependent variable “Duration (F11)”. In case of duration prediction of high rise buildings, the selected features are “Land size (F2)”, “Floor area (F4)”, “No. of basement

(F5)", "Workforce level (F9)" and "No. of floor (F10)" as the p-values of these features are less than 0.05. The rest of the features have not been considered for developing the model.

Table 5.10: Results of one-way ANOVA performed on the collected data of high rise buildings for dependent variable "Duration (F11)".

Feature	**Sum of squares**	**Degree of freedom**	**Mean square**	**F**	**P-value**	**Status**
Location (F1)						
Between Groups	59.810	4	14.952	1.259	0.326	Rejected
Within Groups	190.000	16	11.875			
Total	249.810	20	_			
Land size (F2)						
Between Groups	96.032	1	96.032	11.865	0.003	Selected
Within Groups	153.778	19	8.094			
Total	249.810	20	_			
Floor Height (F3)						
Between Groups	48.595	2	24.298	2.174	0.143	Rejected
Within Groups	201.214	18	11.179			
Total	249.810	20	_			
Floor area (F4)						
Between Groups	99.858	1	99.858	12.653	0.002	Selected
Within Groups	149.952	19	7.892			
Total	249.810	20	_			
No. of basement (F5)						
Between Groups	161.082	1	161.082	34.494	0.000	Selected
Within Groups	88.727	19	4.670			
Total	249.810	20	_			
Design (F6)						
Between Groups	6.495	2	3.248	0.240	0.789	Rejected
Within Groups	243.314	18	13.517			
Total	249.810	20	_			
Finishing materials type (F7)						
Between Groups	14.381	2	7.190	0.550	0.586	Rejected

Within Groups	235.429	18	13.079			
Total	249.810	20	_			
Approval complexity (F8)						
Between Groups	34.476	2	17.238	1.441	0.263	Rejected
Within Groups	215.333	18	11.963			
Total	249.810	20	_			
Workforce level (F9)						
Between Groups	142.117	2	71.059	11.877	0.001	Selected
Within Groups	107.692	18	5.983			
Total	249.810	20	_			
No. of floor (F10)						
Between Groups	88.087	2	44.044	04.902	0.020	Selected
Within Groups	161.722	18	8.902			
Total	249.810	20	_			

Table 5.11 represents the result of one-way ANOVA F-test performed on the collected data for dependent variable “Cost (F11)” of high rise building. In case of cost prediction of high rise buildings, the p-values for “Land size (F2)”, “Floor area (F4)”, “No. of basement (F5)”, “Workforce level (F9)” and “No. of floor (F10)” are less than 0.05 which indicate that these features are statistically significant. On the other hand, p-values for “Location (F1)”, “Floor height (F3)”, “Design (F6)”, “Finishing material type (F7)”, and “Approval complexity (F8)” are larger than 0.05. Therefore, these features were not selected.

Table 5.11: Results of one-way ANOVA performed on the collected data of medium rise buildings for dependent variable “Cost (F12)”.

Feature	**Sum of squares**	**Degree of freedom**	**Mean square**	**F**	**P-value**	**Status**
Location (F1)						
Between Groups	6.883	4	1.721	0.881	0.497	Rejected
Within Groups	31.249	16	1.953			
Total	38.131	20	_			
Land size (F2)						
Between Groups	7.975	1	7.975	5.027	0.037	Selected

Within Groups	30.156	19	1.587			
Total	38.131	20	_			
Floor Height (F3)						
Between Groups	1.846	2	0.923	0.458	0.640	Rejected
Within Groups	36.286	18	2.016			
Total	38.131	20	_			
Floor area (F4)						
Between Groups	18.585	1	18.585	18.066	0.000	Selected
Within Groups	19.546	19	1.029			
Total	38.131	20	_			
No. of basement (F5)						
Between Groups	21.002	1	21.002	23.296	0.000	Selected
Within Groups	17.129	19	0.902			
Total	38.131	20	_			
Design (F6)						
Between Groups	0.422	2	0.211	0.101	0.905	Rejected
Within Groups	37.710	18	2.095			
Total	38.131	20	_			
Finishing materials type (F7)						
Between Groups	1.449	2	0.724	0.355	0.706	Rejected
Within Groups	36.683	18	2.038			
Total	38.131	20	_			
Approval complexity (F8)						
Between Groups	7.119	2	3.560	2.066	0.156	Rejected
Within Groups	31.012	18	1.723			
Total	38.131	20	_			
Workforce level (F9)						
Between Groups	19.962	2	9.981	9.888	0.001	Selected
Within Groups	18.169	18	1.009			
Total	38.131	20	_			
No. of floor (F10)						
Between Groups	23.768	3	7.923	9.377	0.001	Selected

Within Groups	14.364	17	0.845			
Total	38.131	20	_			

Figure 5.4 and 5.6 represent the main effect plot of duration and cost prediction for high rise building, respectively. From the main effect plot, it is seen that the lines for “Land size (F2)”, “Floor area (F4)”, “No. of basement (F5)”, “Workforce level (F9)”, and “No. of floor (F10)” are not horizontal. Therefore, there exists main effect on the response feature for these features. On the other hand, the lines for the rest of the features are almost horizontal indicating no main effect on the response feature.

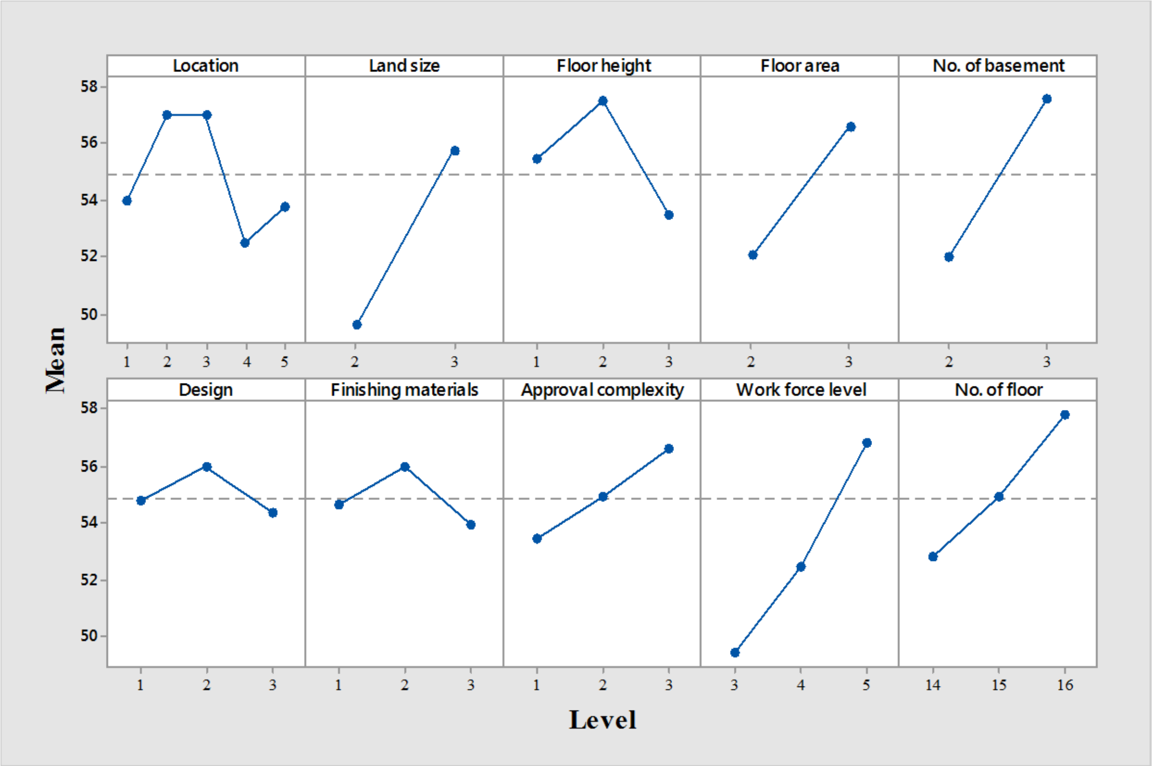


Figure 5.5: Main effect plot for “Duration (F11)” prediction of high rise buildings.

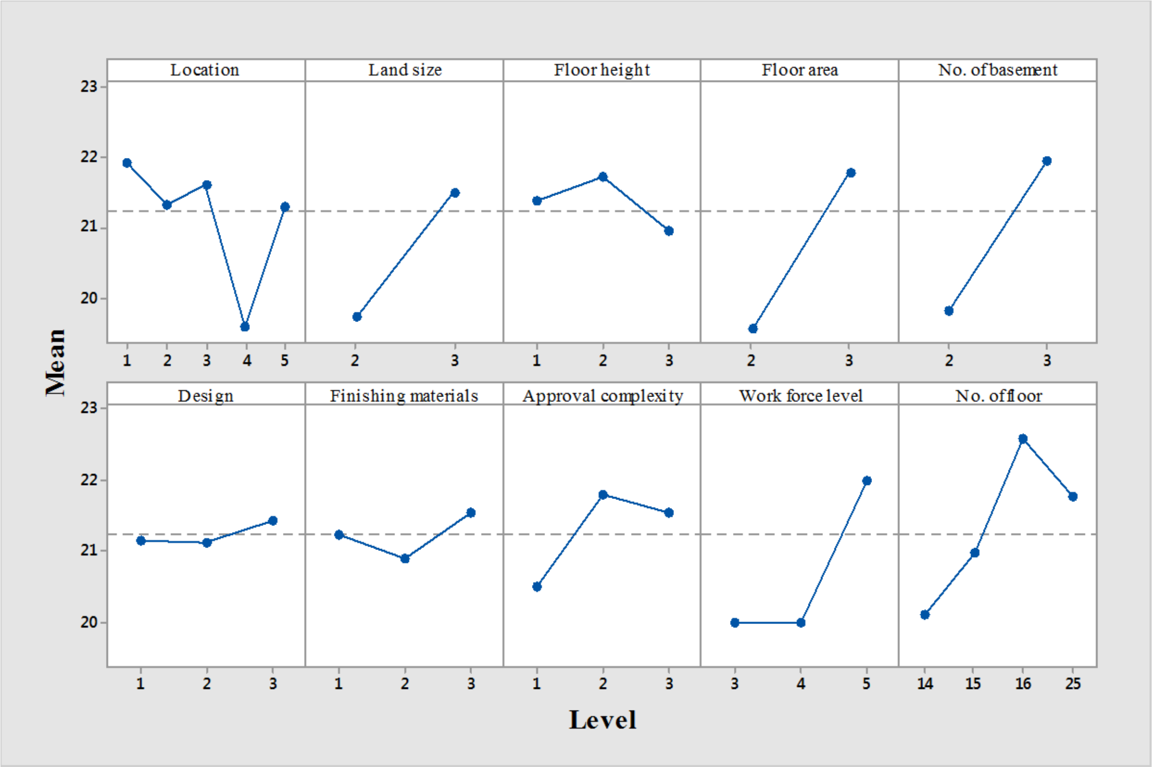


Figure 5.6: Main effect plot for “Cost (F12)” prediction of high rise buildings.

5.5 Building the Model

At this stage, two models were built for construction schedule and cost prediction for each category of building with corresponding features selected in the previous step. Models were built using Python 3.6. Primarily, the collected dataset was divided into two parts. One part of the dataset was for training the models and this is known as training set. Another part of the data set was for testing the models and this is known as testing set. In the current study, training set was constructed by choosing 70% of the total data randomly and rest 30% of the data were considered as testing dataset. Then the values of hyper-parameters of Gradient Boosted Regression Tree (GBRT) for each model were chosen using Random Search method. The hyper-parameter can be defined as the parameters of the model whose values are required to set before training the models.

5.6 Hyper-parameter Regularization using Random Search

Random Search method was applied on the training dataset to select the values of the hyper-parameters for determining the best model performances for each category of building. This process of selecting the values of hyper-parameters is also known as hyper-parameter regularization or hyper-parameter tuning. In the current study, for hyper-parameter tuning of Gradient Boosted Regression Tree (GBRT) schedule and cost prediction models, k-fold cross validation has been used. K-fold cross validation means that the total dataset will be divided into k equal parts. One part of the k parts will be used for training the model. Then, the model is fitted to k-1 parts of the data. The mean r-squared value also known as mean cross validation score has been calculated on the kth part for randomly selected hyper-parameters in random search method. In the current study, the cross validation value of k was set to 3. Therefore, the cross validation score was calculated by taking the average of r-squared values for k=1, 2 and 3.

In order to establish regularized Gradient Boosted Regression Tree (GBRT) models, values of different hyper-parameters have to be identified for best results of the models. There are two categories of hyper-parameters Gradient Boosted Regression Tree (GBRT) algorithm. Learning rate and number of trees are included in the first category of hyper-parameters of GBRT. Learning rate is the step size which is also known as shrinkage parameters. Learning rate can be used to improve the model's generalization capacity.

The first category of hyper-parameter includes learning rate and number of trees. Learning rate is the shrinkage parameter which can be tuned to improve the generalization ability of the model. The number of sub-models in the GBRT is known as number of trees. These two hyper-parameters are used to adjust the gradient boost. The second category of hyper-parameters includes maximum depth and minimum sample leaf. Maximum depth in GBRT is defined as the number of nodes from the longest path. Minimum sample leaf can be defined as the minimum number of leaf nodes.

In this study, random search method has been used for determining the values of gradient boost hyper-parameters to achieve the best estimating effect and improve the model's generalization capacity. The values of hyper-parameters were selected based on the highest mean cross validation r-squared value. This value was considered as the performance metric.

Figure 5.7 represents the mean r-squared values for different number of trees for construction time and cost prediction models of low, medium and high rise buildings. Initially, the number of trees were selected randomly from a normal distribution. Then, random search method has been applied. From figure, it is seen that based on the highest mean cross validation r-squared values, the number of trees for duration prediction models of low, medium and high rise buildings are 100. On the other hand, the number of trees for cost prediction models of low rise building is 75, and for medium and high rise building, this value is 125.

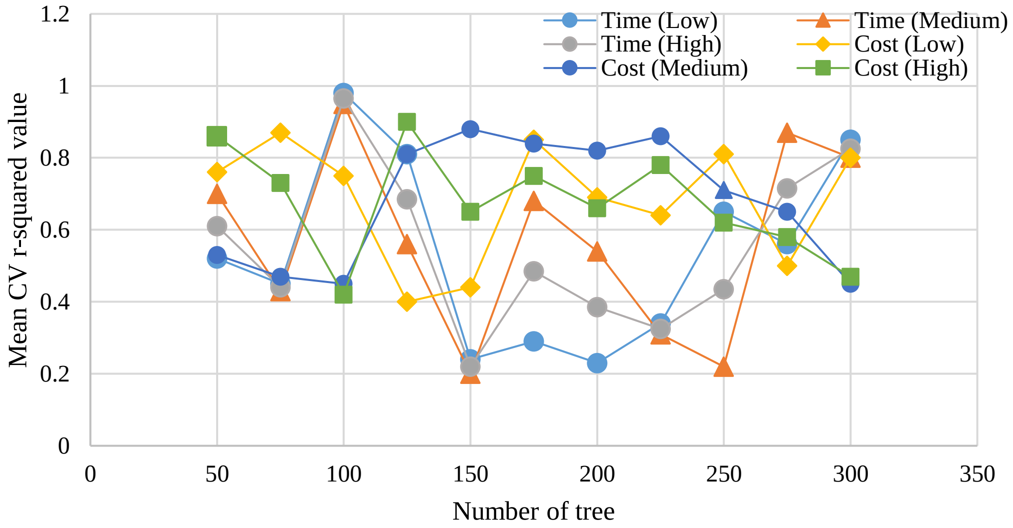


Figure 5.7: Mean CV r-squared values for different number of trees.

Figure 5.8 represents the mean r-squared values for different number of sample leaf for construction time and cost prediction model. Initially, the number of sample leaf were selected randomly from a normal distribution. Then, random search method has been applied. Based on the highest mean cross validation r-squared values, the number of sample leaf for duration prediction models of low, and medium rise buildings is 5 and for high rise building, the value is 4. On the other hand, the numbers of sample leaf for cost prediction models of low, medium and high are 5, 4 and 6.

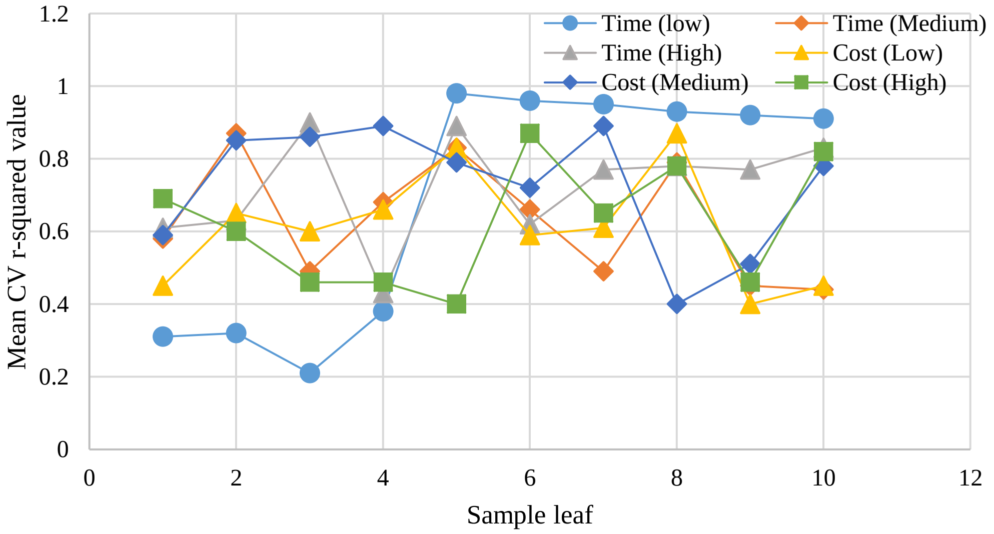


Figure 5.8: Mean CV r-squared values for different number of sample leaf.

Figure 5.9 represents the mean r-squared values for different depth of trees for construction duration and cost prediction model. Initially, different numbers of depth of trees were selected randomly from a normal distribution. Then, random search method has been applied. Based on the highest mean cross validation r-squared values, the number of depth of trees for duration prediction models of low, medium and high rise buildings are 3, 7 and 6. On the other hand, the numbers of depth of trees for cost prediction models of low rise building is 6, and for medium and high rise building, this value is 5.

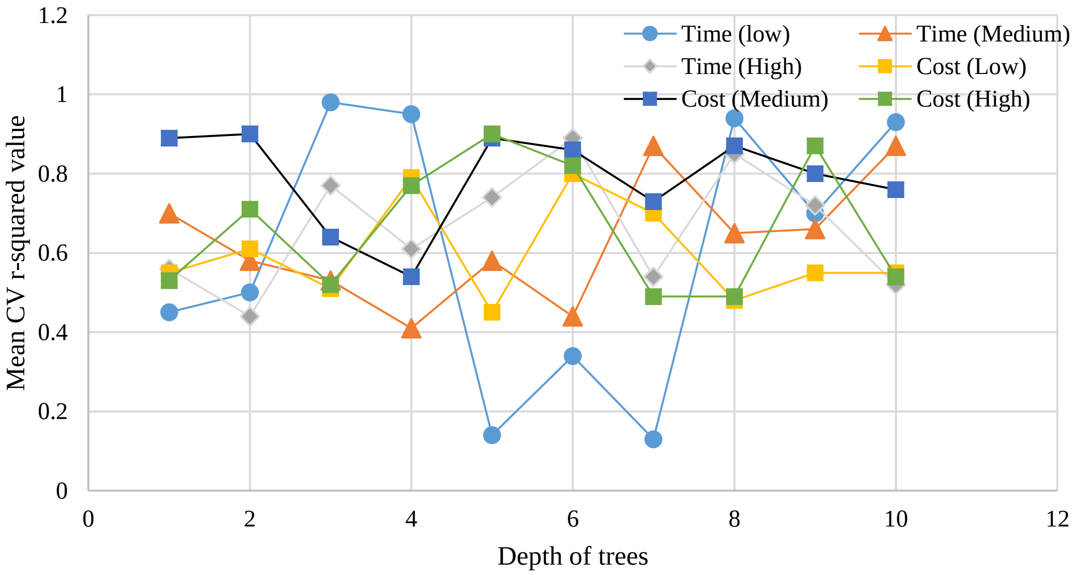


Figure 5.9: Mean CV r-squared values for different depth of trees.

Figure 5.10 represents the mean r-squared values for different learning rates for construction duration and cost prediction model. Initially, different learning rates were selected randomly from a normal distribution. Then, random search method has been applied. Based on the highest mean cross validation r-squared values, the learning rates for duration prediction models of low, medium and high rise buildings are 0.6, 0.1 and 0.6. On the other hand, learning rate for cost prediction models of low and medium rise building is 0.4, and for medium, this value is 0.5.

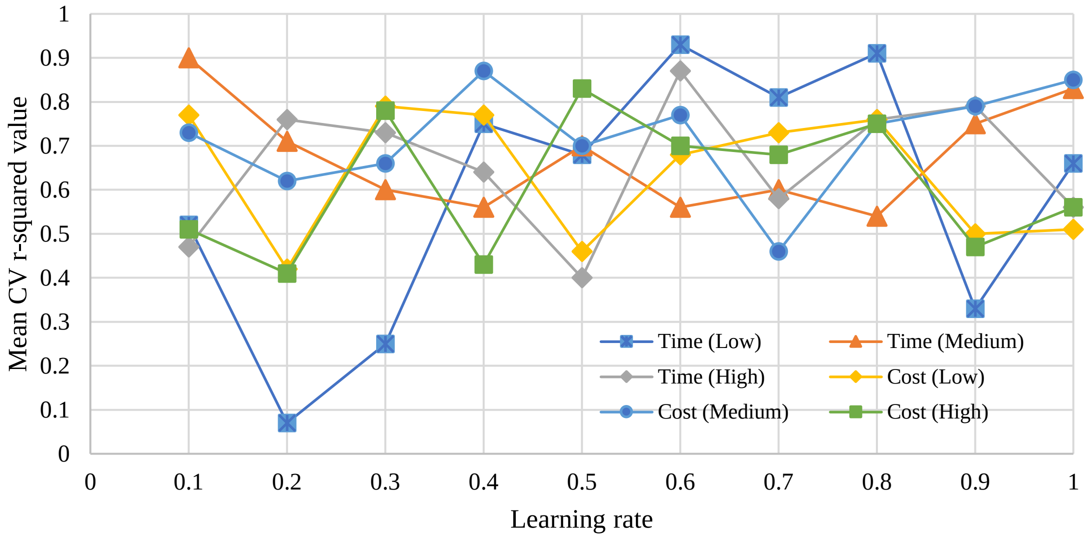


Figure 5.10: Mean CV r-squared values for different learning rates.

Table 5.12 shows the summary of selected values of the four hyper-parameters of Gradient Boosted Regression Tree (GBRT) for developing construction schedule and cost prediction models for low, medium and high rise buildings.

Table 5.12: Selected values of the hyper-parameters for the models.

	Number of trees	**Minimum sample leaf**	**Maximum depth**	**Learning rate**
Time (low rise)	100	5	3	0.6
Time (medium rise)	100	5	7	0.1
Time (high rise)	100	4	6	0.6
Cost (low rise)	75	4	6	0.4
Cost (medium rise)	125	5	5	0.4
Cost (high rise)	125	6	5	0.5

CHAPTER 6: RESULT AND ANALYSIS

In this section, the performances of the developed models were evaluated. The estimated construction durations and costs of low, medium and high rise buildings based on regularized Gradient Boosted Regression Tree models were not only analyzed against the performance metrics but also compared with the performance results of Support Vector Regression (SVR) and Multiple Linear Regression (MLR) models.

6.1 Performance Evaluation of the Models

Several papers have discussed about the performance evaluation metrics for regression models **[Belavagi & Muniyal, 2016; Strom et al., 2019]**. Some of these metrics are mean absolute percentage error (MAPE), r-squared value, mean squared error (MSE), etc. In order to evaluate the predictive performance, regularized Gradient Boosted Regression Tree (GBRT) based construction schedule and cost prediction models were measured by r-squared value, mean absolute percentage error (MAPE) and mean squared error (MSE).

The comparison between the actual and predicted construction duration of low rise buildings of the regularized Gradient Boosted Regression Tree (GBRT) model for training dataset has been shown in Figure 6.1. Here, the blue rectangle stands for the true or actual value of construction time duration while the orange triangle indicates the predicted construction duration for low rise buildings. It is noted that there is no significant difference between actual and predicted duration for the training model.

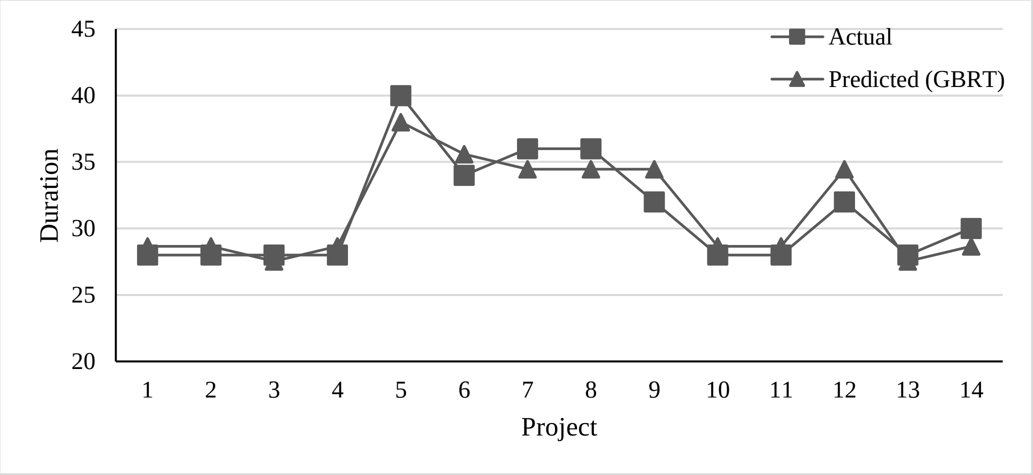


Figure 6.1: Actual and predicted duration of the regularized GBRT schedule prediction model of low rise building for training data.

For the testing data, the comparison between the actual and predicted construction durations of low rise buildings of the regularized Gradient Boosted Regression Tree (GBRT) model has been shown in Figure 6.2. Here, the blue rectangle stands for the true or actual value of construction time duration while the orange triangle indicates the predicted construction duration for low rise buildings. It is noted that there is also no significant difference between actual and predicted duration for the testing model. To check the performance more precisely, r-squared value and mean absolute percentage error (MAPE) have been calculated.

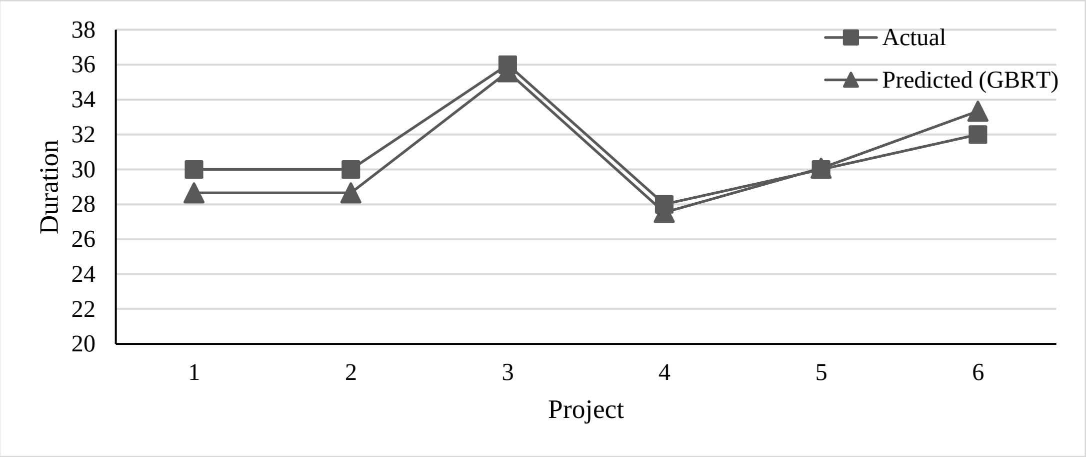


Figure 6.2: Actual and predicted duration of the regularized GBRT schedule prediction model of low rise building for testing data.

The r-squared value of the regularized Gradient Boosted schedule prediction model for training data of low rise building is shown in Figure 6.3. The r-squared value is 0.79 which indicates very good predictive performance for the training data.

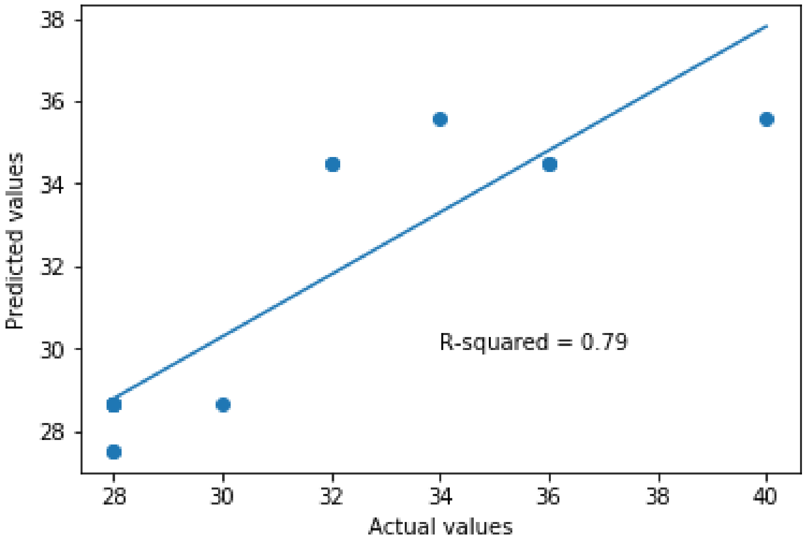


Figure 6.3: R-squared value of the regularized GBRT schedule prediction model of low rise building for training data.

The training set and test set deviance (mean squared error) of GBRT model for low rise building has been shown in Figure 6.4. The deviance has been expressed as the mean squared error which is the function of the number of iterations for the regularized Gradient Boosted Regression Tree based construction schedule prediction model. The optimal value of the boosting iteration was set at the point for which the deviance (mean squared error) was minimum. Here, it is noted that the boosting iteration value is 120 where the test set deviance tends to decrease and the mean squared value is 0.9715 for the testing set. The deviance between training set and test set is very low. This is happened because GBRT has the ability to minimize the performance gap between training and test set.

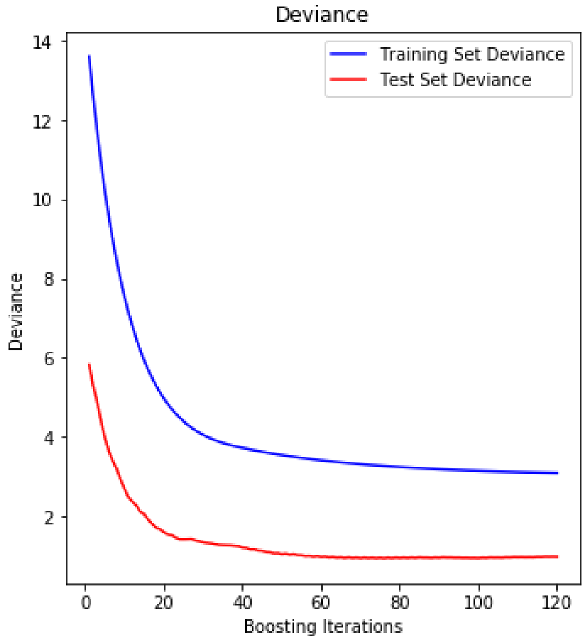


Figure 6.4: Testing set deviance of the regularized GBRT schedule prediction model for low rise building.

The comparison between the actual and predicted construction duration of medium rise buildings of the regularized Gradient Boosted Regression Tree (GBRT) model for training dataset has been shown in Figure 6.5. Here, the blue rectangle stands for the true or actual value of construction time duration of medium rise buildings while the orange triangle indicates the predicted construction duration. It is noted that for medium rise building, there is no significant difference between actual and predicted duration for the training model.

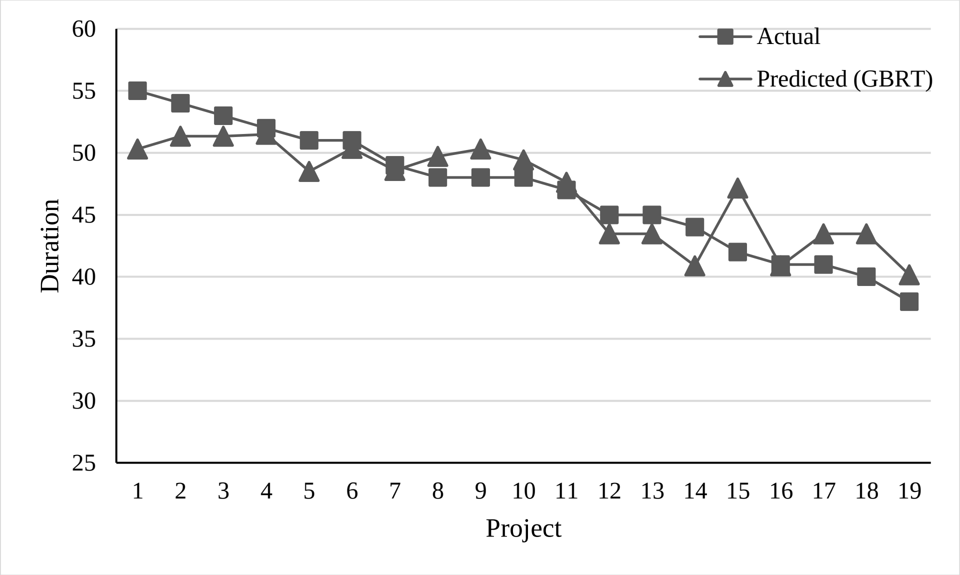


Figure 6.5: Actual and predicted duration of the regularized GBRT schedule prediction model of medium rise building for training data.

For the testing dataset, the comparison between the actual and predicted construction duration of medium rise buildings of the regularized Gradient Boosted Regression Tree (GBRT) model has been shown in Figure 6.6. In case of testing model, it is noted that for medium rise building, there is little difference between actual and predicted duration for some projects while no significant differences have been found for most of the projects.

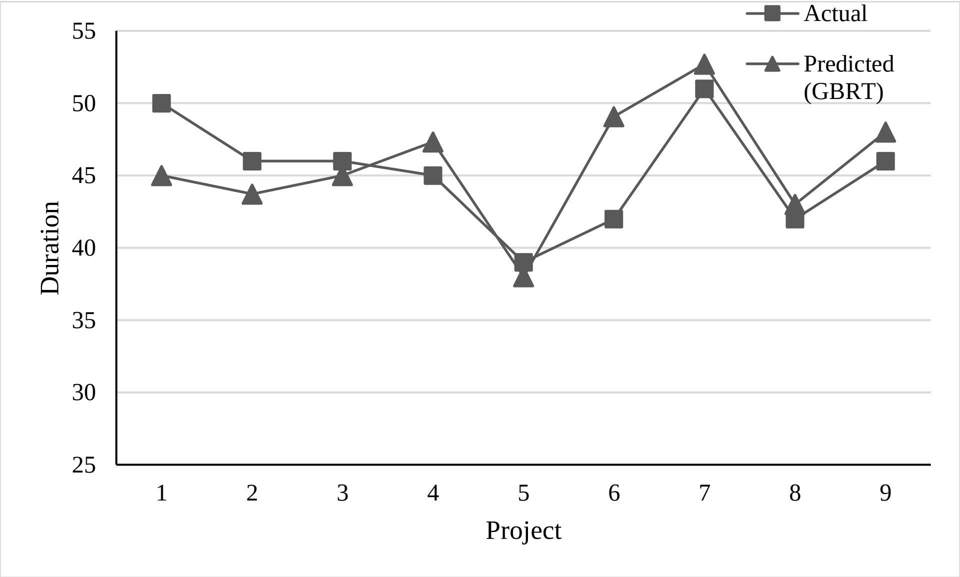


Figure 6.6: Actual and predicted duration of the regularized GBRT schedule prediction model of medium rise building for testing data.

The r-squared value of the regularized Gradient Boosted schedule prediction model for training data of medium rise building is shown in Figure 6.7. The r-squared value is 0.75 which indicates very good predictive performance for the training data.

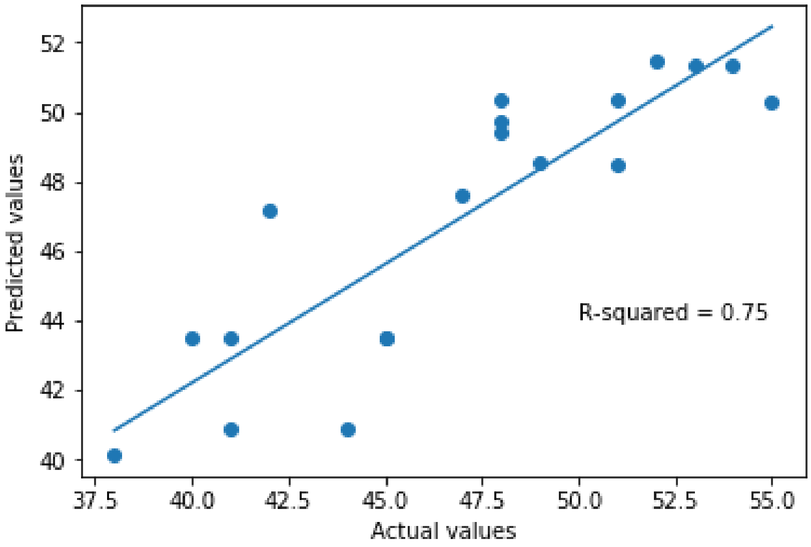


Figure 6.7: R-squared value of the regularized GBRT schedule prediction model of medium rise building for training data.

The training set and test set deviance (mean squared error) of GBRT model for medium rise building has been shown in Figure 6.8. Here, it is noted that the boosting iteration value is 100 where the test set deviance tends to decrease and the mean squared value is 20.0078 for the testing set. Here, the deviance between training set and test set higher than the low rise building.

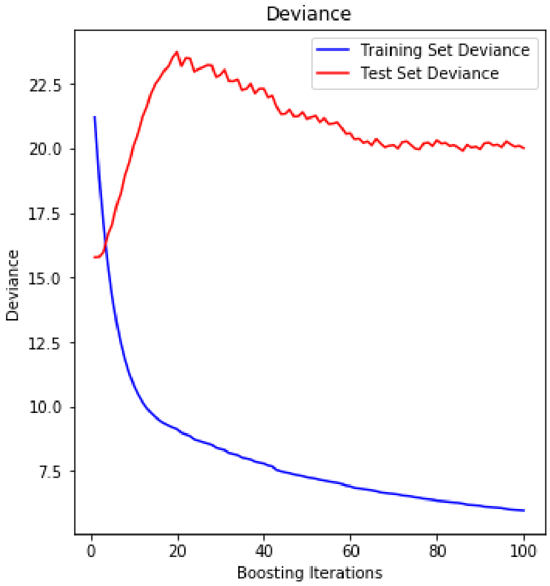


Figure 6.8: Testing set deviance of the regularized GBRT schedule prediction model for medium rise building.

Figure 6.9 shows the comparison between the actual and predicted construction duration of high rise buildings of the regularized Gradient Boosted Regression Tree (GBRT) model for training dataset. Here, the blue rectangle stands for the true or actual value of construction time duration of high rise buildings while the orange triangle indicates the predicted construction duration. It is noted that there is no significant difference between actual and predicted duration.

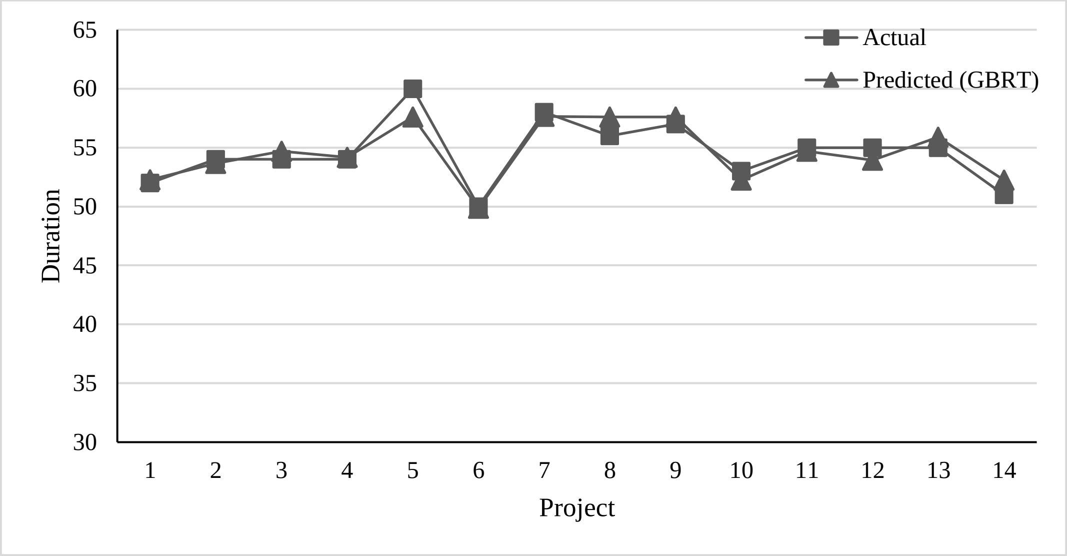


Figure 6.9: Actual and predicted duration of the regularized GBRT schedule prediction model of high rise building for training data.

Figure 6.10 shows the comparison between the actual and predicted duration in case of high rise buildings for testing data. It is noted that there is no significant difference between actual and predicted duration. This indicates the very satisfactory performance of GBRT model.

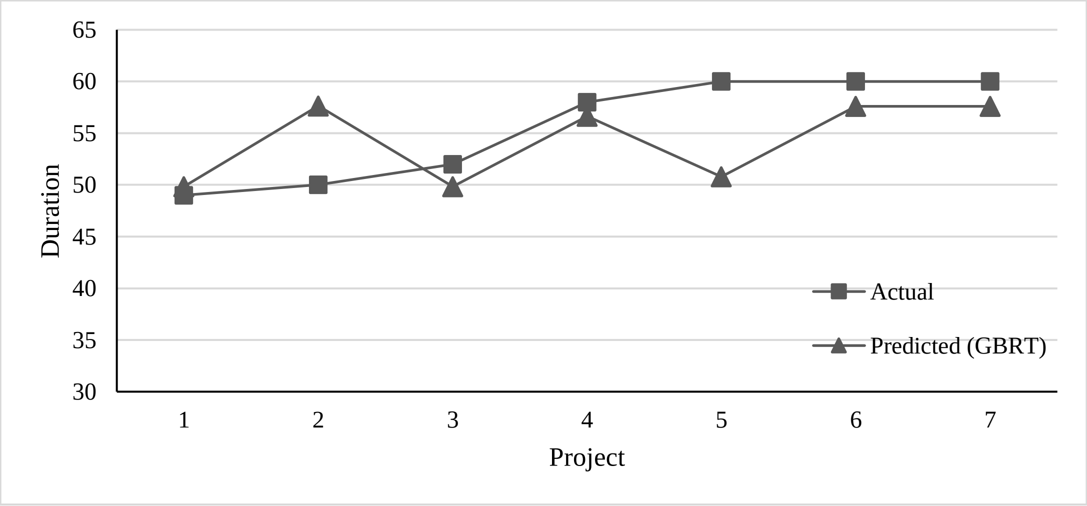


Figure 6.10: Actual and predicted duration of the regularized GBRT schedule prediction model of high rise building for testing data.

The r-squared value of the regularized Gradient Boosted schedule prediction model for training data of high rise building is shown in Figure 6.11. The r-squared value is 0.85 which indicates very good predictive performance for the training data in case of high rise buildings.

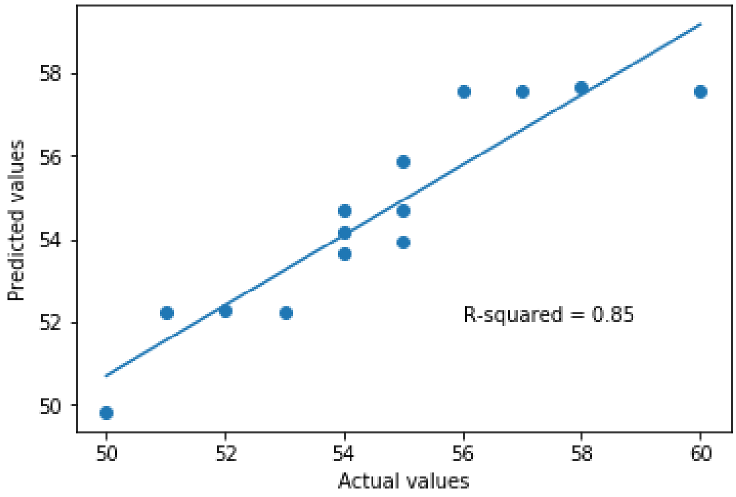


Figure 6.11: R-squared value of the regularized GBRT schedule prediction model of high rise building for training data.

Figure 6.12 represents the training set and test set deviance (mean squared error) of GBRT model for high rise buildings. Here, it is noted that the boosting iteration value is 100 where the test set deviance tends to decrease and the mean squared value is 23.2011 for the testing set. For high rise building, the training and test set deviance is higher than low and medium rise building. As the number of floor of the building increases, it becomes very difficult to predict the construction schedule.

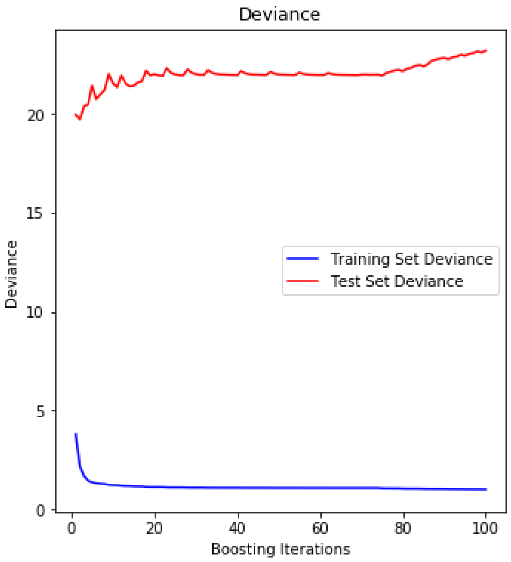


Figure 6.12: Testing set deviance of the regularized GBRT schedule prediction model for high rise building.

The comparison between the actual and predicted construction cost of low rise buildings of the regularized Gradient Boosted Regression Tree (GBRT) model for training dataset has been shown in Figure 6.13. Here, the blue rectangle stands for the true or actual value of construction cost of medium rise buildings while the orange triangle indicates the predicted construction costs. It is noted that for low rise building, there is no significant difference between actual and predicted costs for the training model.

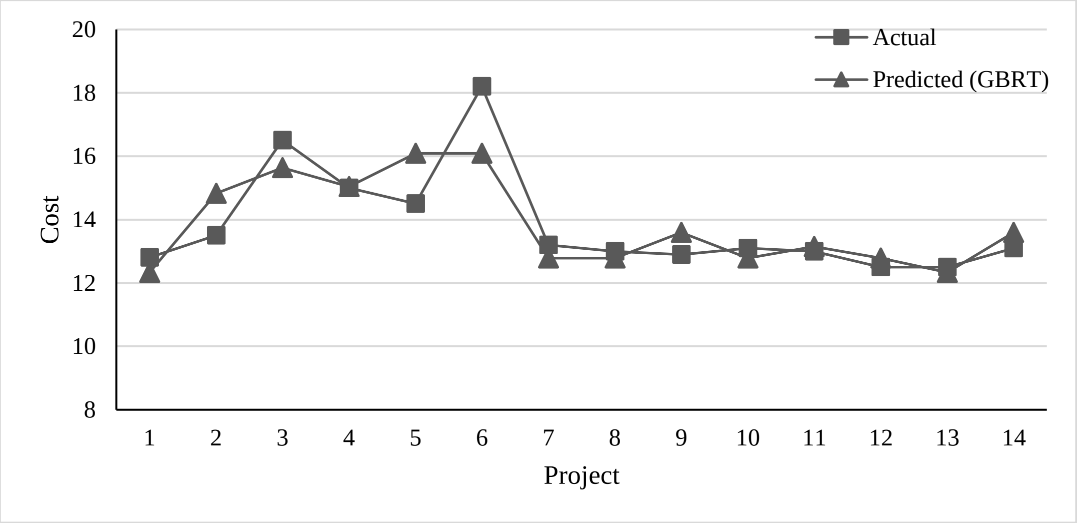


Figure 6.13: Actual and predicted cost of the regularized GBRT cost prediction model of low rise building for training data.

Figure 6.14 shows the comparison between the actual and predicted costs in case of high rise buildings for testing data. It is noted that there is no significant difference between actual and predicted cost.

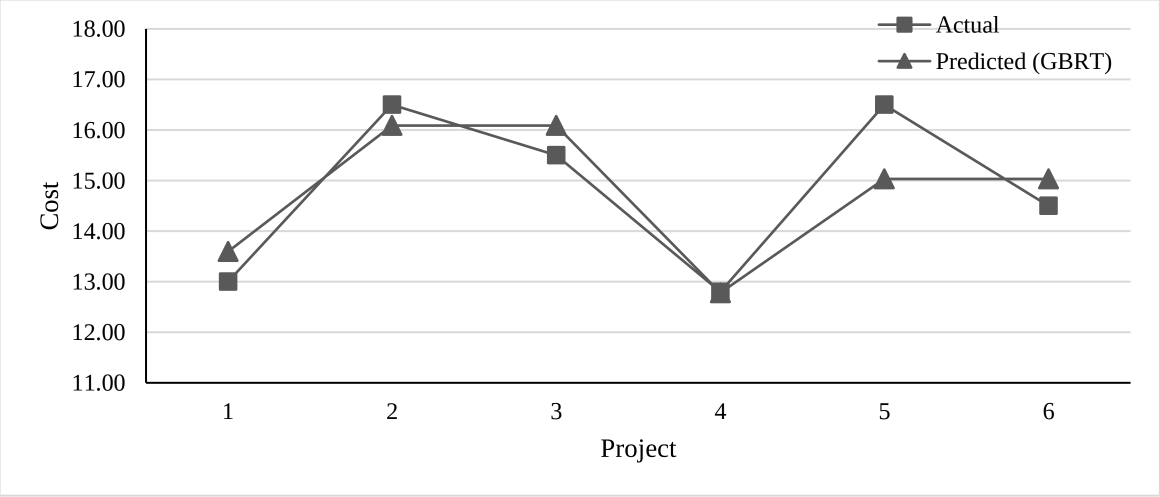


Figure 6.14: Actual and predicted cost of the regularized GBRT cost prediction model of low rise building for testing data.

Figure 6.15 shows the r-squared value of the regularized Gradient Boosted cost prediction model for training data of low rise building. The r-squared value is 0.70 which indicates very good predictive performance because higher r-squared value indicates satisfactory performance.

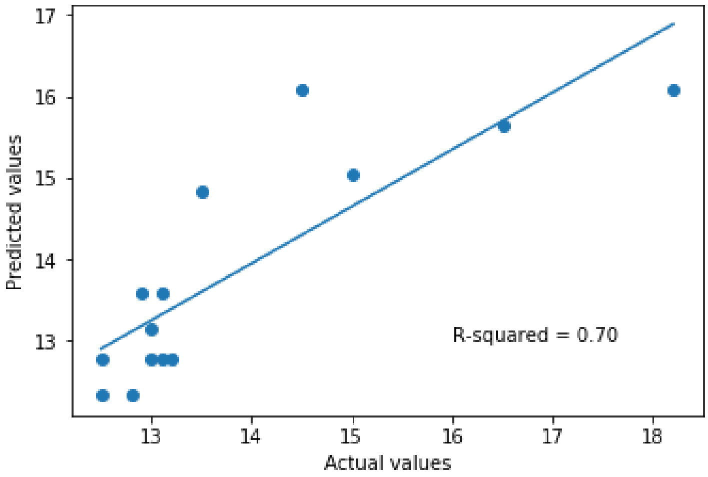


Figure 6.15: R-squared value of the regularized GBRT cost prediction model of low rise building for training data.

Figure 6.16 represents the training set and test set deviance (mean squared error) of GBRT cost prediction model for low rise buildings. Here, it is noted that the boosting iteration value is 70 where the test set deviance tends to decrease and the mean squared value is 0.5507 for the testing set. There is no significant difference between training set and test set deviance. Therefore, it can be said that the overfitting problem is minimized in this case. The deviance is minimized because GBRT has the ability to reduce overfitting problems.

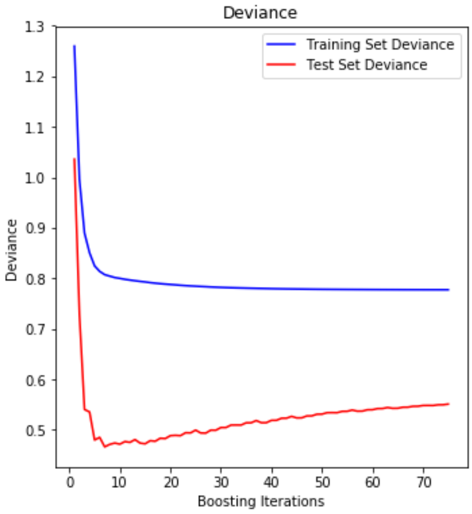


Figure 6.16: Testing set deviance of the regularized GBRT cost prediction model for low rise building.

Figure 6.17 shows the comparison between the actual and predicted construction cost of medium rise buildings of the regularized Gradient Boosted Regression Tree (GBRT) model for training dataset. Here, the blue rectangle stands for the true or actual value of construction cost of medium rise buildings while the orange triangle indicates the predicted construction costs. It is noted that for medium rise building, there is no significant difference for the training model.

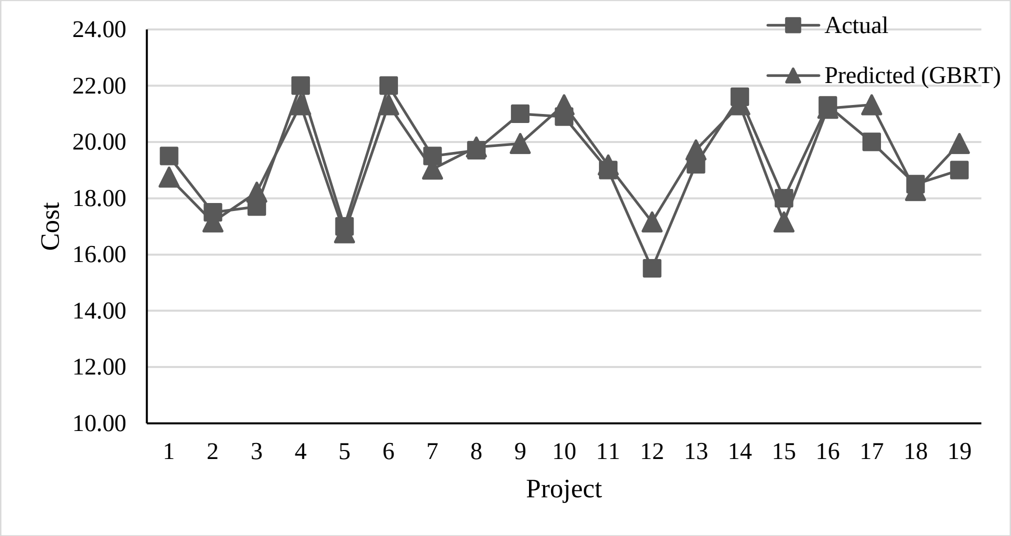


Figure 6.17: Actual and predicted cost of the regularized GBRT cost prediction model of medium rise building for training data.

Figure 6.18 shows the comparison between the actual and predicted costs of the GBRT model in case of medium rise buildings for testing data. It is noted that there exists little difference between actual and predicted cost.

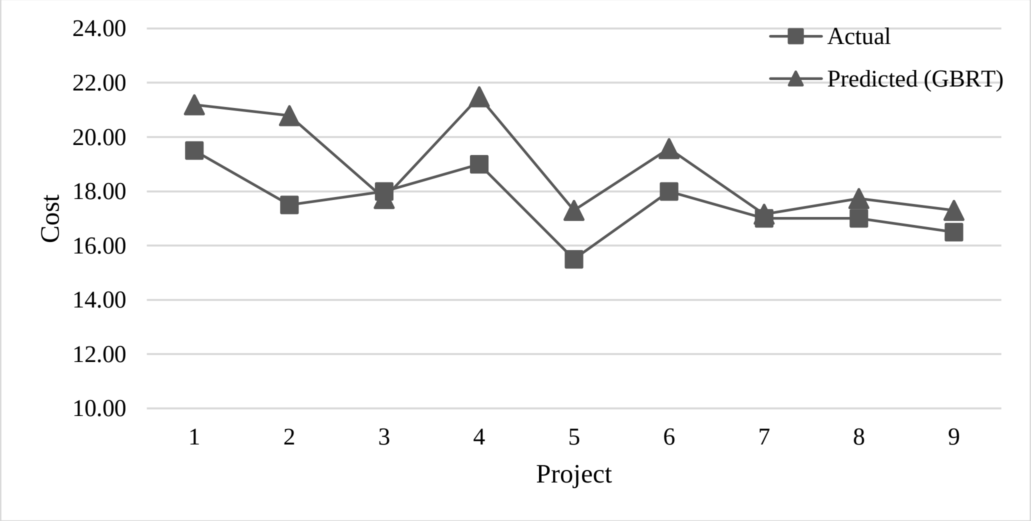


Figure 6.18: Actual and predicted cost of the regularized GBRT cost prediction model of medium rise building for testing data.

The r-squared value of the regularized Gradient Boosted cost prediction model for training data of medium rise building is shown in Figure 6.19. The r-squared value is 0.83 which indicates very good predictive performance for the training data in case of medium rise buildings.

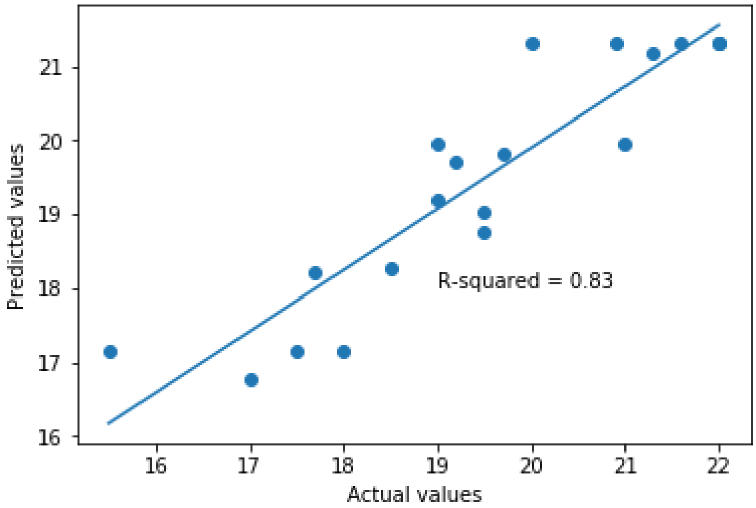


Figure 6.19: R-squared value of the regularized GBRT cost prediction model of medium rise building for training data.

Figure 6.20 represents the training set and test set deviance (mean squared error) of GBRT cost prediction model for medium rise buildings. Here, it is noted that the boosting iteration value is 120 where the test set deviance tends to decrease and the mean squared value is 2.9834 for the testing set. Here, the deviance is higher than the low rise building. It is difficult to predict the construction cost of medium rise buildings than low rise buildings.

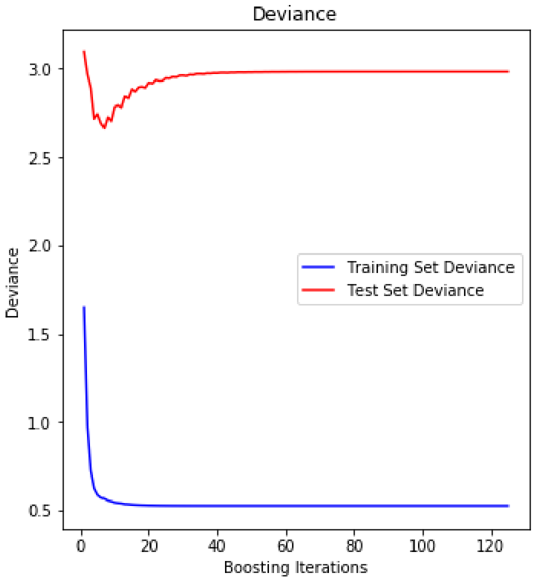


Figure 6.20: Testing set deviance of the regularized GBRT cost prediction model for medium rise building.

Figure 6.21 shows the comparison between the actual and predicted construction cost of high rise buildings of the regularized Gradient Boosted Regression Tree (GBRT) model for training dataset. Here, the blue rectangle stands for the true or actual value of construction cost of medium rise buildings while the orange triangle indicates the predicted construction costs. It is noted that for high rise building, there is no significant difference between the actual and predicted costs for the training model.

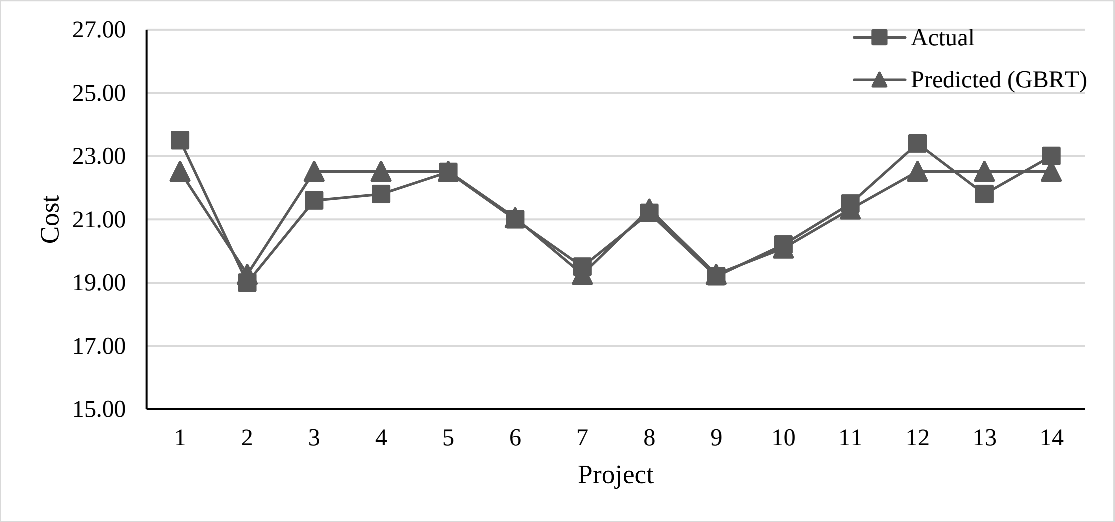


Figure 6.21: Actual and predicted cost of the regularized GBRT cost prediction model of high rise building for training data.

Figure 6.22 shows the comparison between the actual and predicted costs of the GBRT model in case of high rise buildings for testing data. It is noted that there is no significant difference between actual and predicted cost in case of testing model.

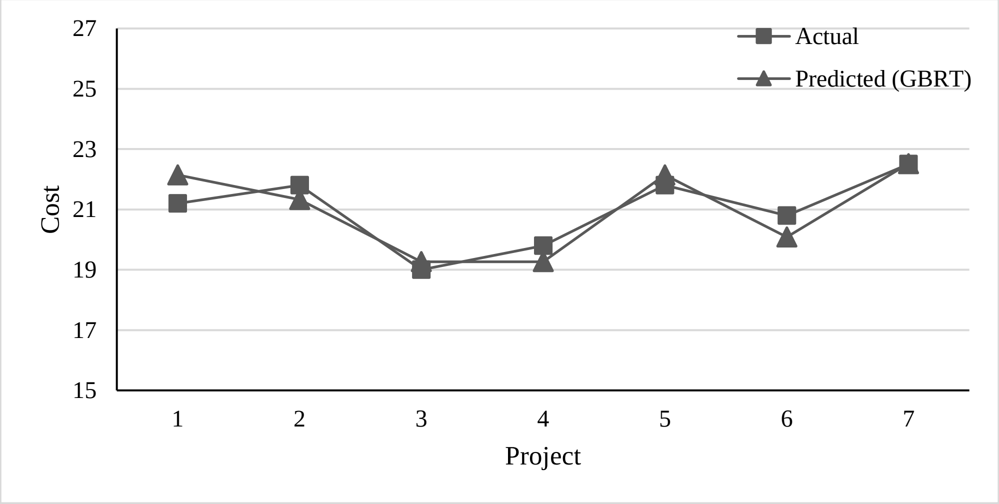


Figure 6.22: Actual and predicted cost of the regularized GBRT cost prediction model of high rise building for testing data.

The r-squared value of the regularized Gradient Boosted cost prediction model for training data of high rise building is shown in Figure 6.23. The r-squared value is 0.86 which indicates very good predictive performance for the training data in case of high rise buildings.

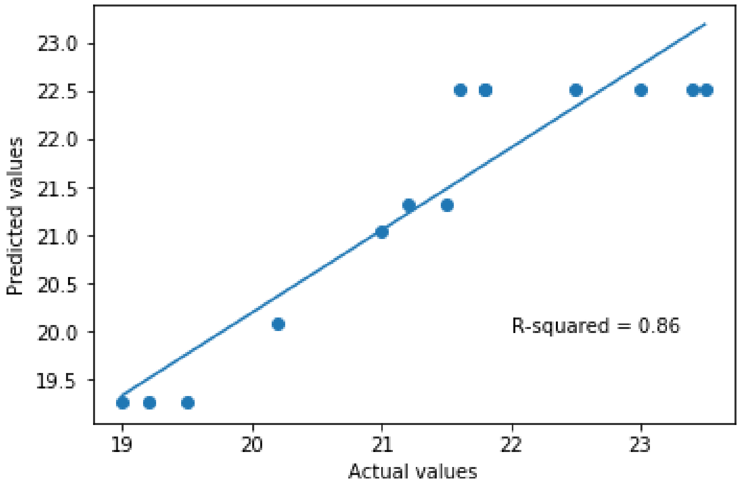


Figure 6.23: R-squared value of the regularized GBRT cost prediction model of high rise building for training data.

Figure 6.24 represents the training set and test set deviance (mean squared error) of GBRT cost prediction model for high rise buildings. Here, it is noted that the boosting iteration value is 70 where the test set deviance tends to decrease and the mean squared value is 0.3003 for the testing set.

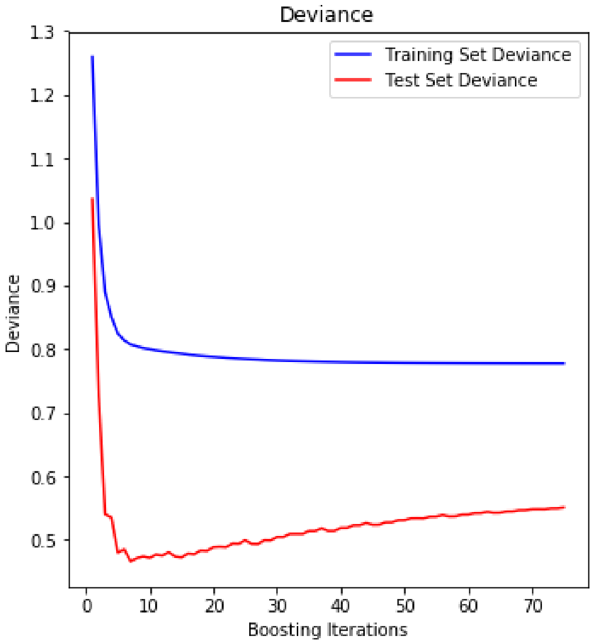


Figure 6.24: Testing set deviance of the regularized GBRT cost prediction model for high rise building.

6.2 Feature importance

One of the advantages of Gradient Boosted Regression Tree (GBRT) model is that interpretation of features can be done based on importance. Applying GBRT, the ranking of features can be obtained based on their contributions to the model performance. Usually features are used to split branches in GBRT. The feature which gives relatively low training error than other feature while splitting the branch are ranked better than that feature.

Figure 6.25, Figure 6.26 and Figure 6.27 display the relative importance of the most influential predictor variables of the construction schedule prediction model for low rise, medium rise and high rise building respectively. Since these measures are relative, total values of these features are equal to 1. From the figures, it is noted that for low rise buildings, “Location (F1)” is the most influential feature since it has got the value of 0.342. Various locations have different working time restrictions. Again, for the medium rise buildings, “No. of floor (F10)” is the most influential feature. In most of the cases, higher the number of floor in the building, higher the duration required for construction. In case of high rise buildings, No. of floor (F10)” is the most influential feature and it has got the value of 0.555.

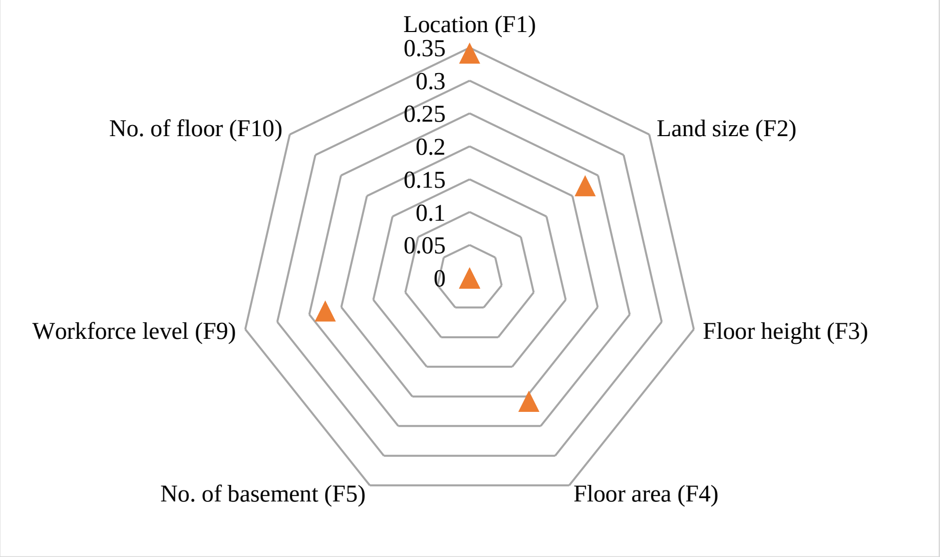


Figure 6.25: Feature importance for the regularized GBRT schedule prediction model low rise building.

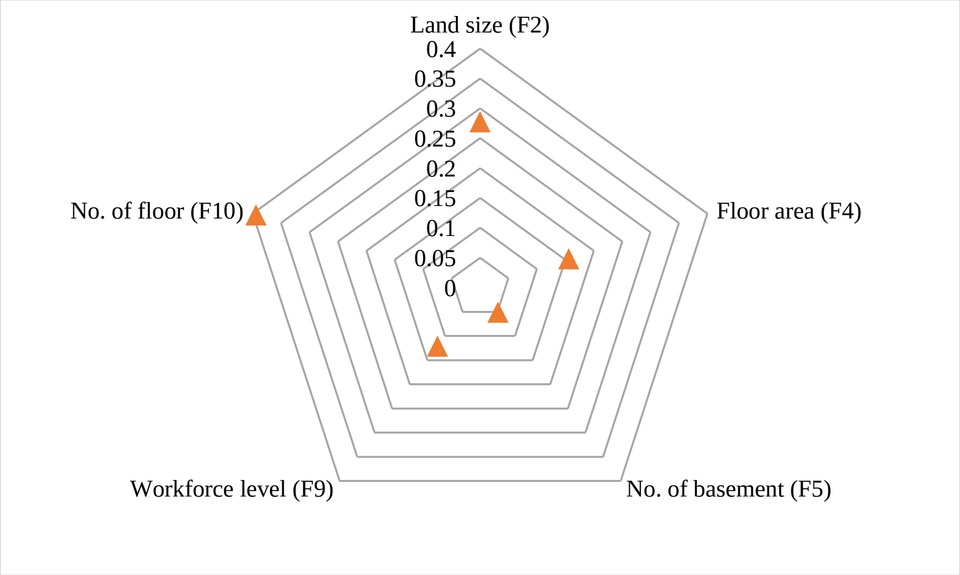


Figure 6.26: Feature importance for the regularized GBRT schedule prediction model medium rise building.

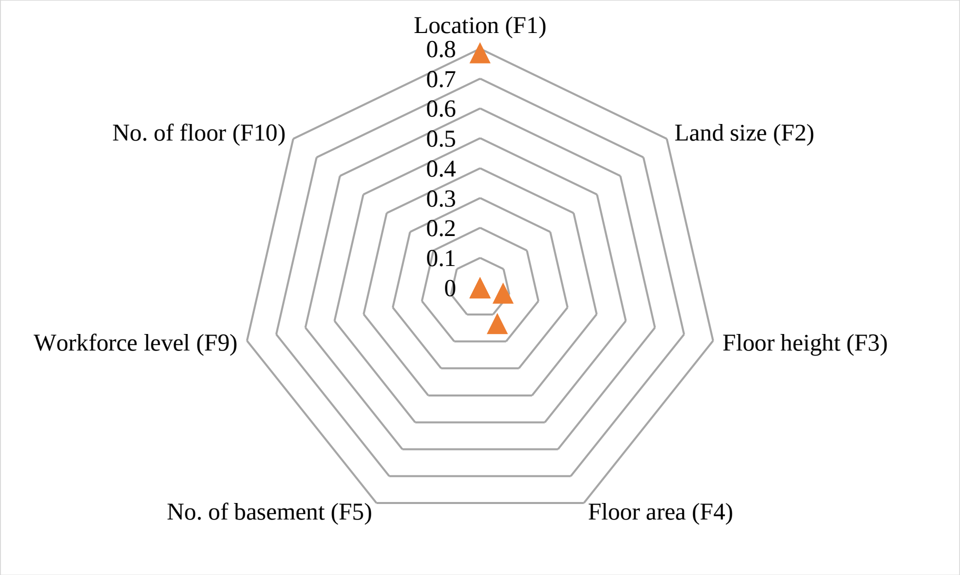


Figure 6.27: Feature importance for the regularized GBRT schedule prediction model high rise building.

Figure 6.28, Figure 6.29 and Figure 6.30 show the relative importance of the most influential predictor variables of the construction cost prediction model for low rise, medium rise and high rise building respectively. From the figures, it is noted that for low rise buildings, “Location (F1)” is the most influential feature for cost prediction since it has got the value of 0.0.787. Various locations have different demands for features and amenities. Again, for the medium rise buildings, “No. of floor (F10)” is the most influential feature and it has got the value of

0.372. In most of the cases, higher the number of floor in the building, higher the duration required for construction. In case of high rise buildings, No. of floor (F10)" is the most influential feature and it has got the value of 0.56.

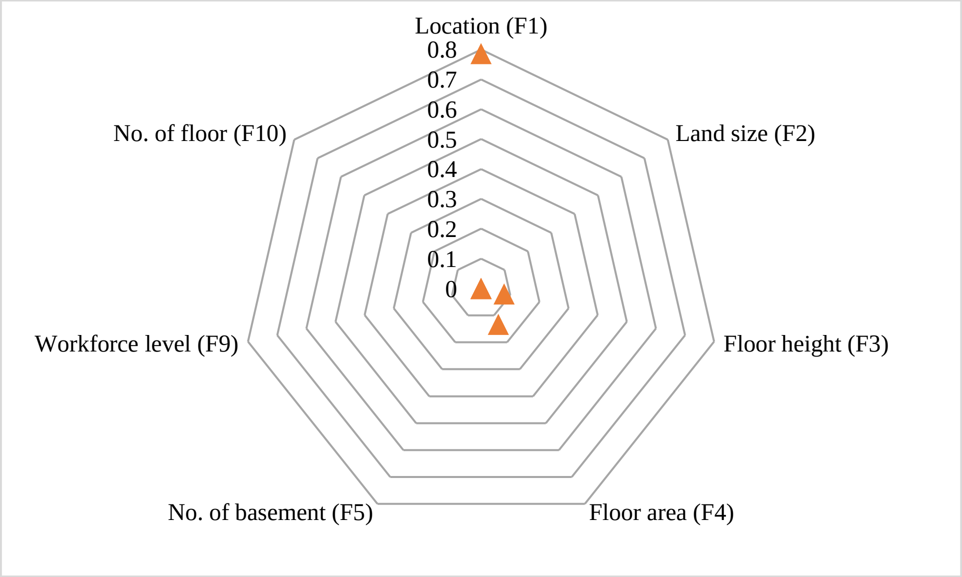


Figure 6.28: Feature importance for the regularized GBRT cost prediction model of low rise building.

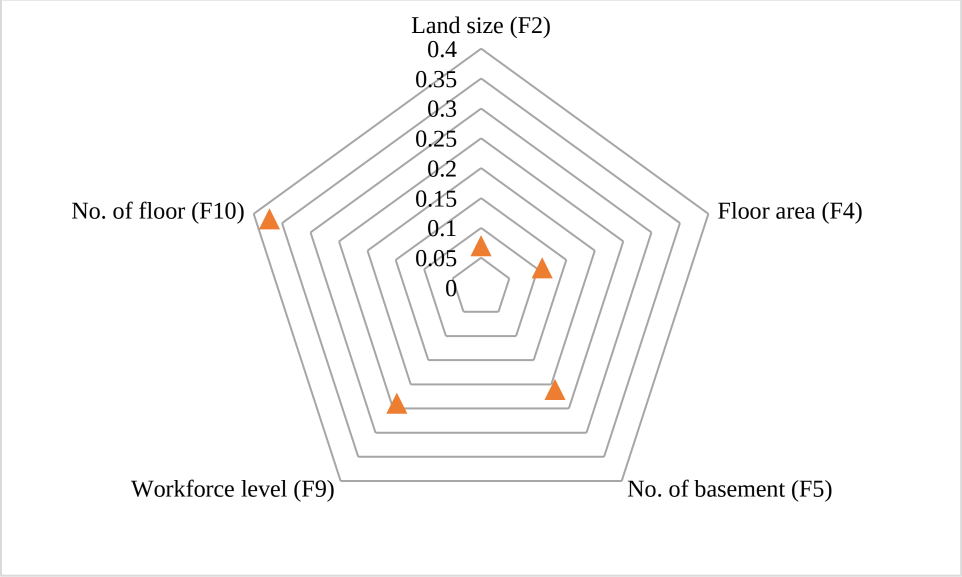


Figure 6.29: Feature importance for the regularized GBRT cost prediction model of medium rise building.

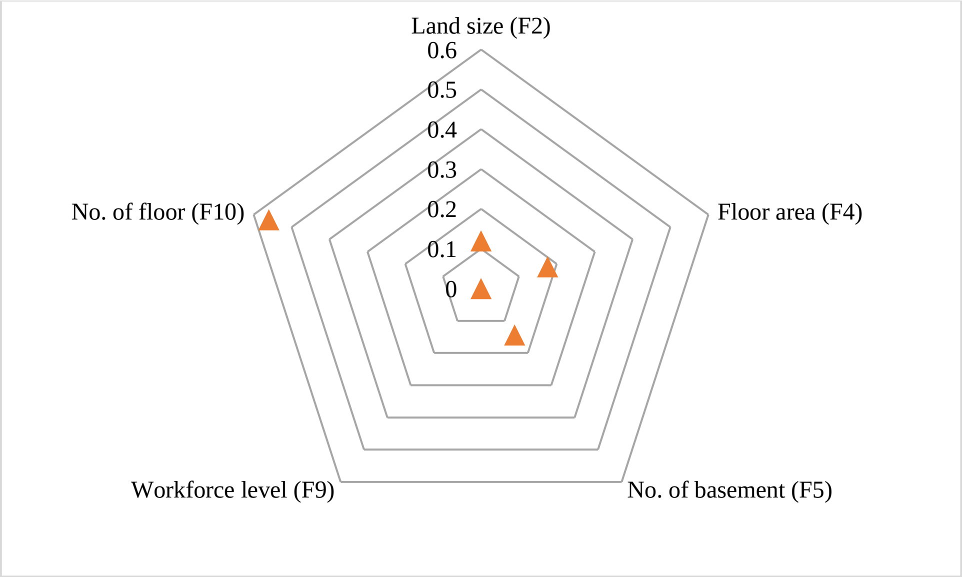


Figure 6.30: Feature importance for the regularized GBRT cost prediction model of high rise building.

6.3 Partial dependence of features

Partial dependence plots show the dependence between the dependent feature and other input features. Additional insights about how the set of input features affect the dependent feature in each model have been represented through the partial dependence plots. Figure 6.31 shows the partial dependence plots of construction schedule prediction model for low rise buildings. The vertical scale is in the log odds and the hash marks on the x-axis represent the deciles of the distribution of the corresponding feature. Partial dependence plots show that “Location (F1)” has a moderate partial dependence impact on the construction duration for low rise buildings. Construction duration has significant partial dependence on “Land size (F2)” and “Floor area (F4)”. Construction duration has not significant partial dependence on “Floor area (F4)” and “No. of basement (F5)”. They show the linear relationship with construction duration since it is positively associated with construction duration. “Location (F1)” has also been identified as the most influential feature of construction duration prediction model for low rise building. Construction duration has also significant partial dependence on “Workforce level (F9)” for low rise buildings. In reality, the workforce level has much impact on the construction time.

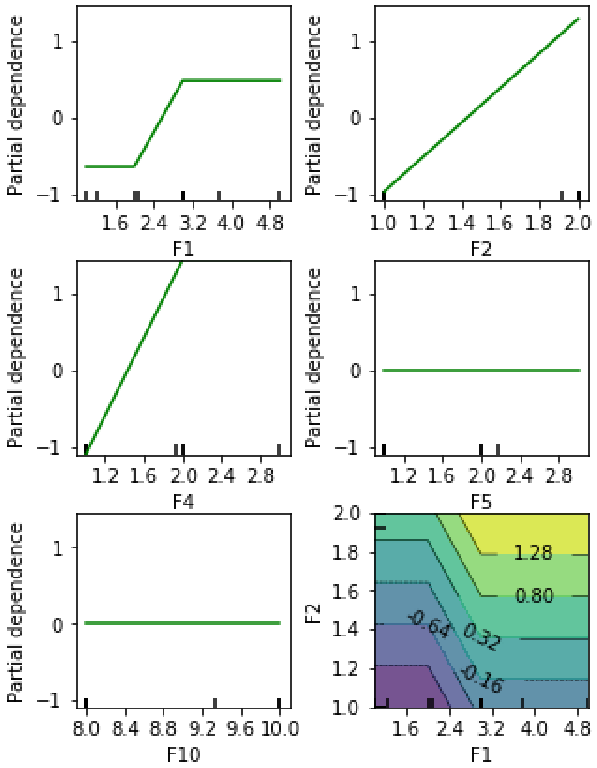
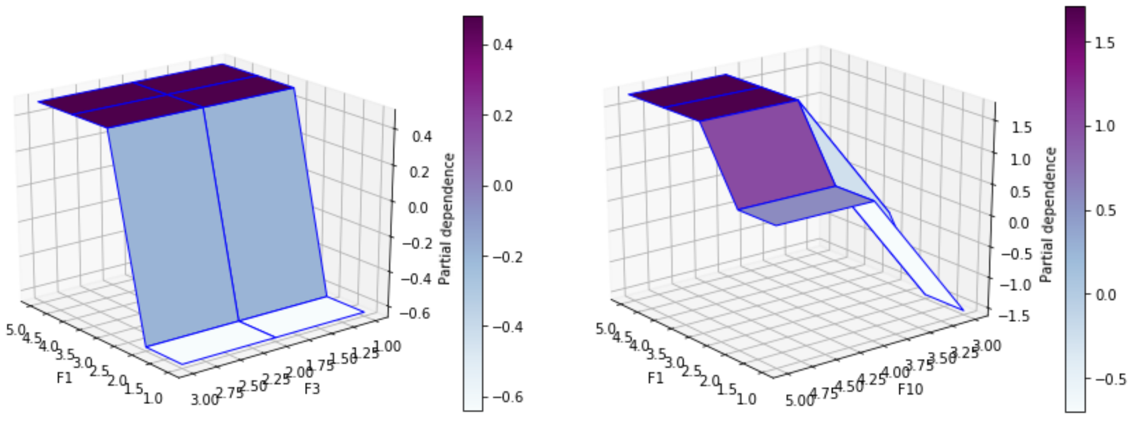


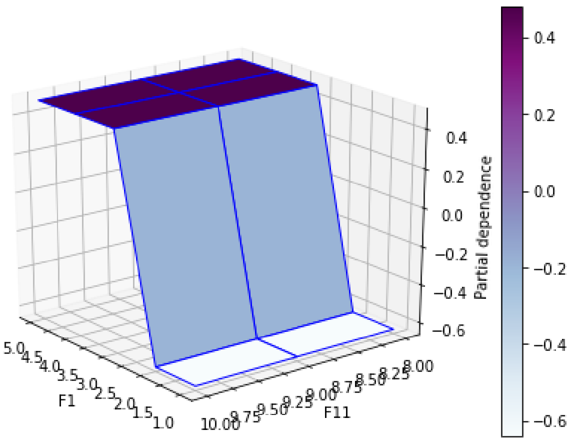
Figure 6.31: Partial dependence plots of features for the regularized GBRT schedule prediction model of low rise building.

An interesting relationship given in Figure 6.32 which shows the joint dependence between "Location (F1)" and other significant factors on construction duration. There appears an interaction effect among these features. Construction duration tends to be higher for type 4 and type 5 of "Location (F1)".



(A)

(B)



(C)

Figure 6.32: Joint partial dependence plots of regularized GBRT schedule prediction model of low rise building.

Figure 6.33 shows the partial dependence plots of construction schedule prediction model for medium rise buildings. Partial dependence plots show that "Land size (F2)", "Floor area (F4)", "No. of basement (F5)" have moderate partial dependence impact on the construction duration for medium rise buildings. Construction duration has significant partial dependence on "Workforce level (F9)" and "No. of floor (F10)".

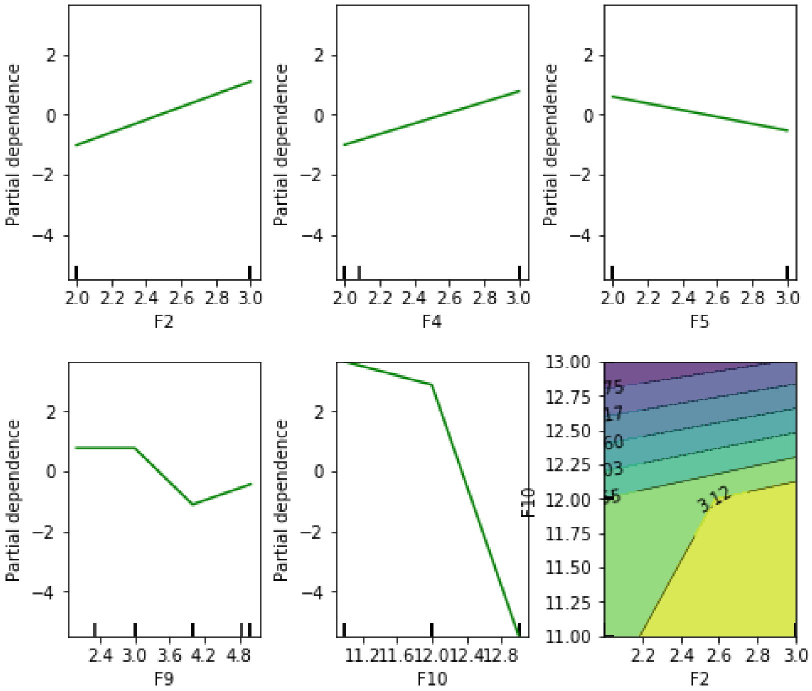
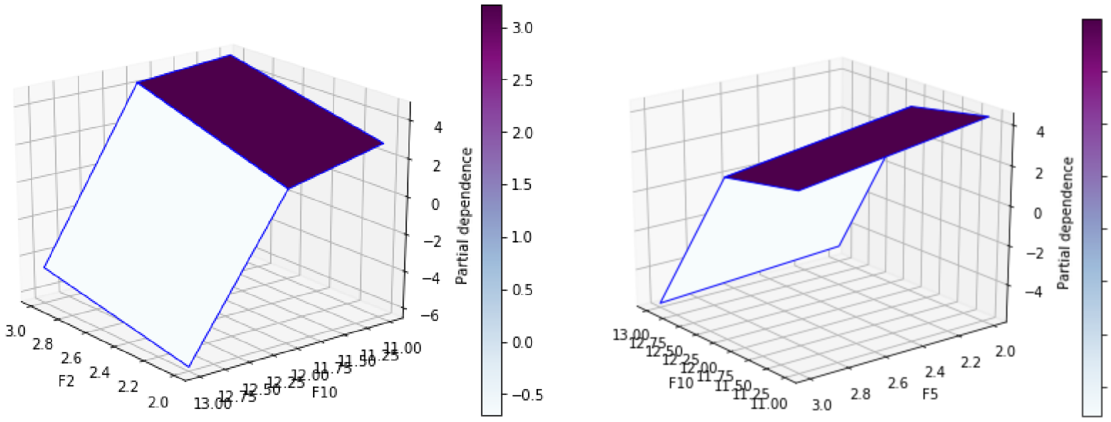


Figure 6.33: Partial dependence plots of features for the regularized GBRT schedule prediction model of medium rise building.

Figure 6.34 which shows the joint dependence between “Land size (F2)”, “No. of basement (F5)” and “No. of floor (F10)” on construction duration for medium rise building. There appears an interaction effect among these features. Construction duration tends to be higher for higher “No. of floor (F10)”.



(A)

(B)

Figure 6.34: Joint partial dependence plots of for regularized GBRT schedule prediction model of medium rise building.

Figure 6.35 shows the partial dependence plots of construction schedule prediction model for high rise buildings. Partial dependence plots show that “Floor area (F4)”, “No. of basement (F5)” have moderate partial dependence impact on the construction duration for high rise buildings. Construction duration has significant partial dependence on “Workforce level (F9)” and “No. of floor (F10)” in case of high rise buildings.

Figure 6.36 which shows the joint dependence between “Workforce level (F9)” and “No. of floor (F10)” on construction duration for medium rise building. There appears an interaction effect among these features. Construction duration tends to be higher for higher “No. of floor (F10)” and lower “Workforce level (F9)”.

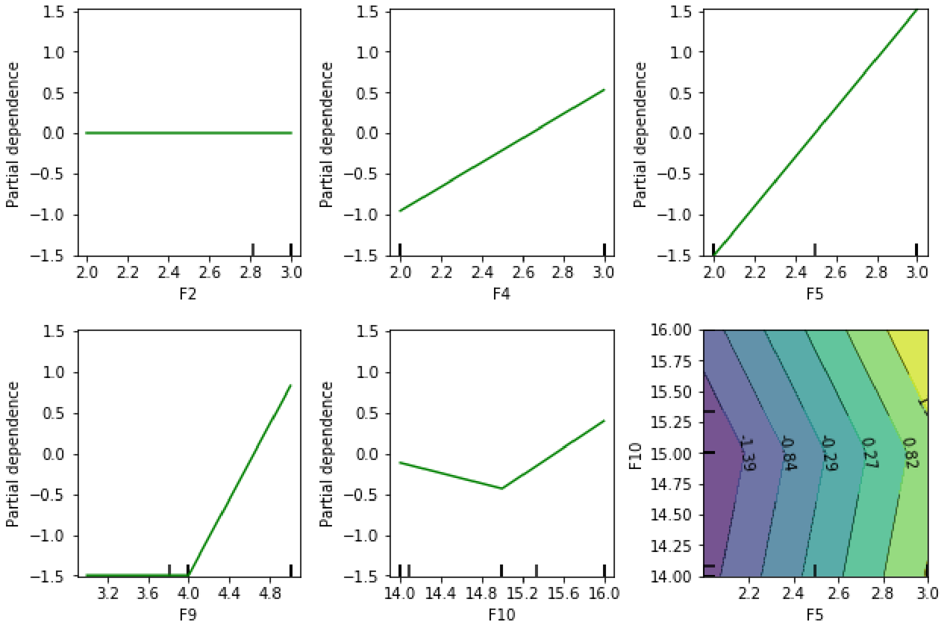


Figure 6.35: Partial dependence plots of features for the regularized GBRT schedule prediction model of high rise building.

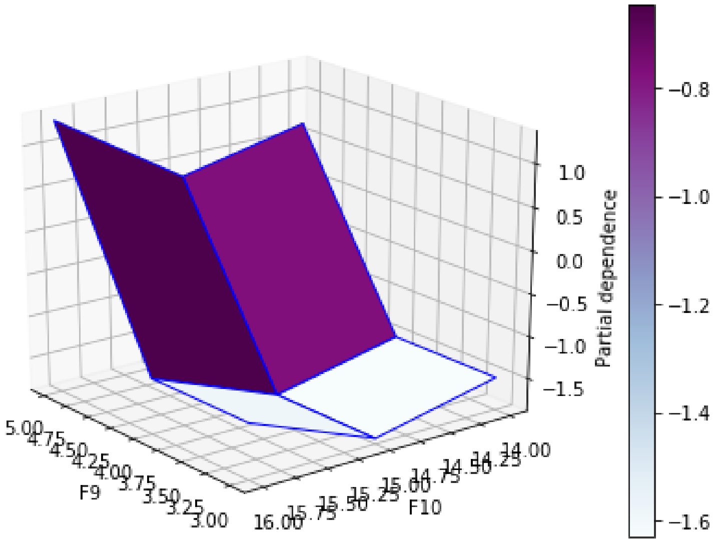


Figure 6.36: Joint partial dependence plots of features with respect to feature F5 for regularized GBRT schedule prediction model of high rise building.

Figure 6.37 shows the partial dependence plots of construction cost prediction model for low rise buildings. Partial dependence plots show that “Location (F1)”, “Floor height (F3)” and “Floor area (F4)” have moderate partial dependence impact on the construction cost for low rise buildings.

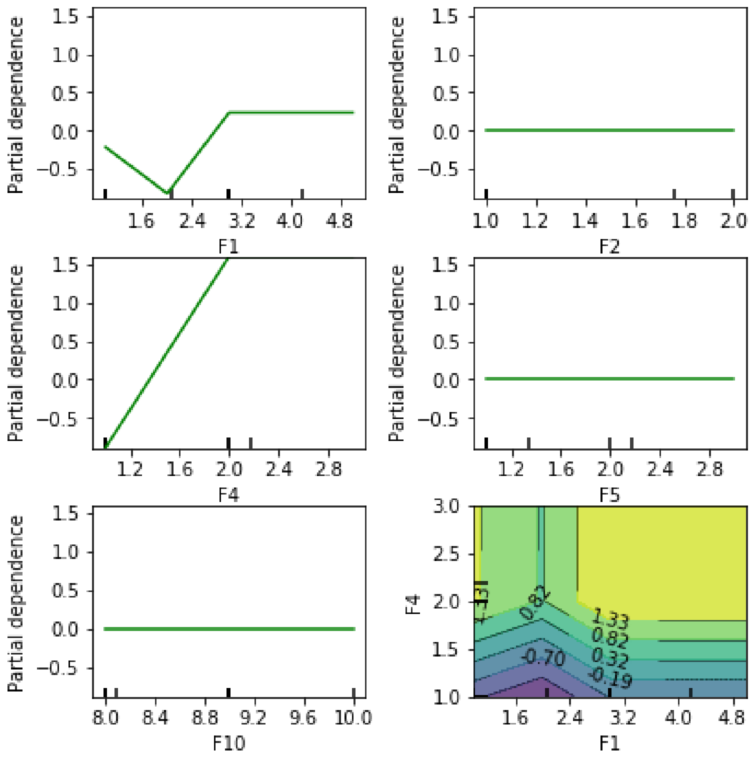
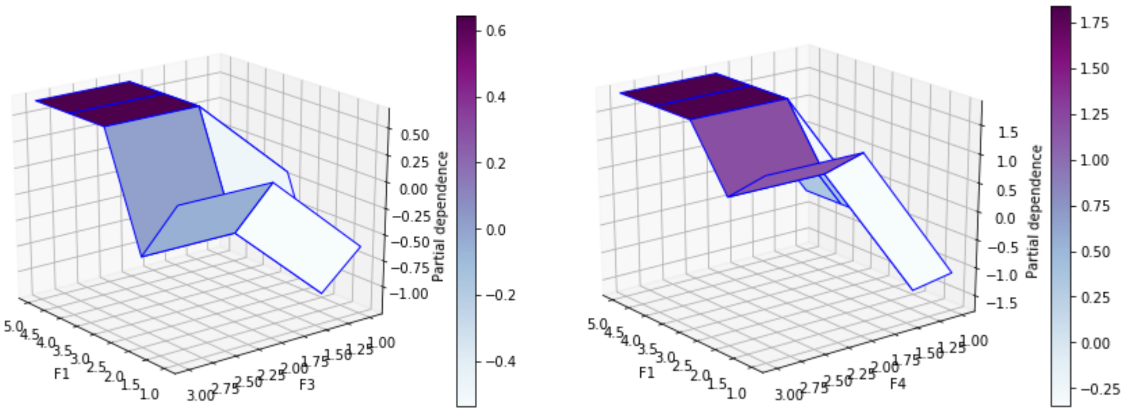


Figure 6.37: Partial dependence plots of features for the regularized GBRT cost prediction model of low rise building.

Figure 6.38 shows the joint dependence among “Location (F1)”, “Floor height (F3)” and “Floor area (F4)” on construction cost. Partial dependence is very high for type 4 and type 5 of “Location (F1)”. Partial dependence is also large for higher “Floor height (F3)” and “Floor area (F4)” for low rise building.



(A)

(B)

Figure 6.38: Partial dependence plots of features for the regularized GBRT cost prediction model of low rise building.

Figure 6.39 shows the partial dependence plots of construction cost prediction model for medium rise buildings. Partial dependence plots show that “Land size (F2)”, “Workforce level (F9)” and “No. of floor (F10)” have significant partial dependence impact on the construction cost for medium rise buildings. “Floor area (F4)” and “No. of basement (F5)” have moderate partial dependence impact on construction cost.

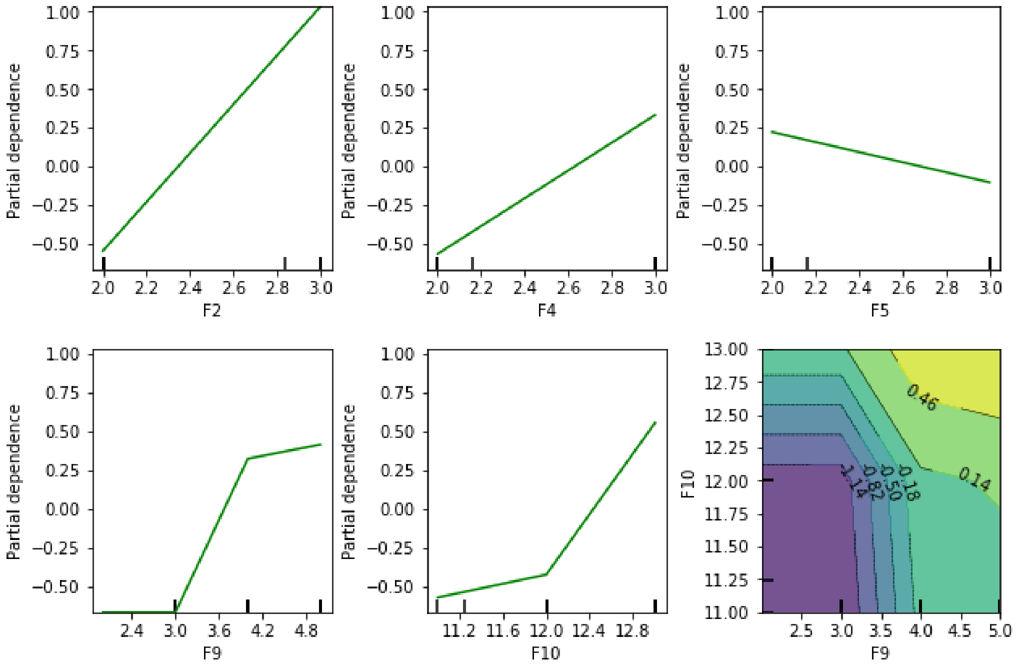


Figure 6.39: Partial dependence plots of features for the regularized GBRT cost prediction model of medium rise building.

Figure 6.40 represents the joint dependence between “Workforce level (F9)” and “No. of floor (F10)” on construction cost.

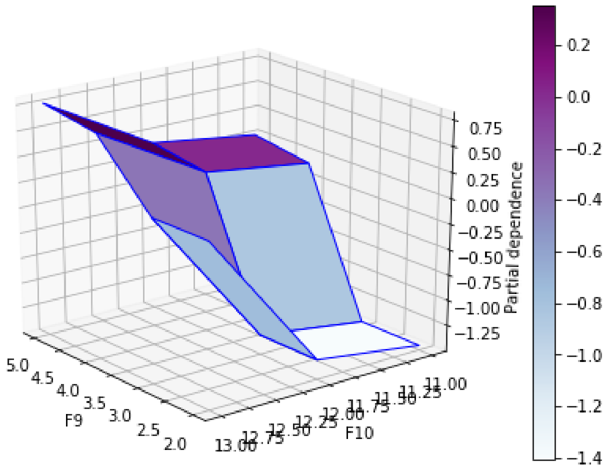


Figure 6.40: Joint partial dependence plots of features for the regularized GBRT cost prediction model of medium rise building.

The partial dependence plots of construction cost prediction model for high rise buildings has been shown in Figure 6.41. Partial dependence plots show that “Floor area (F4)”, “No. of basements (F5)” and “No. of floor (F10)” have significant partial dependence impact on the construction cost for high rise buildings. “Floor height (F2)” has moderate partial dependence impact on construction cost for high rise building.

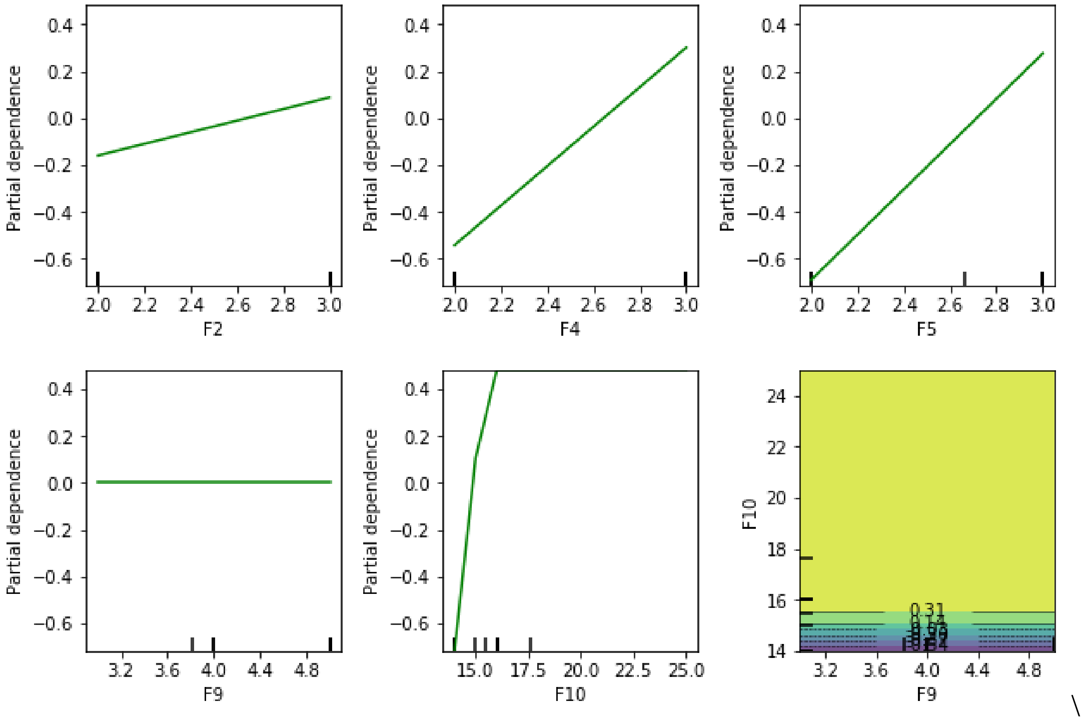


Figure 6.41: Partial dependence plots of features for the regularized GBRT cost prediction model of high rise building.

Figure 6.42 represents the joint dependence between “Workforce level (F9)” and “No. of floor (F10)” on construction cost.

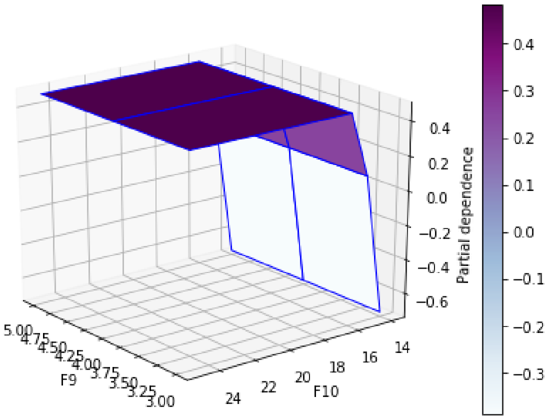


Figure 6.42: Joint partial dependence plots of features for the regularized GBRT cost prediction model of high rise building.

6.4 Performance comparison

In the next step of this study, performances of the regularized Gradient Boosted regression tree (GBRT) models have been compared against Support Vector Regression (SVR) and Multiple Linear Regression (MLR) models based on the test dataset. This was done by comparing the mean absolute percentage error (MAPE) and mean squared error (MSE) of the models. Figure 6.43, Figure 6.44 and Figure 6.45 show the absolute percentage errors (MAPE) for the testing data of low rise, medium rise and high rise buildings for regularized GBRT, Support Vector Regression and Multiple Linear Regression schedule prediction models.

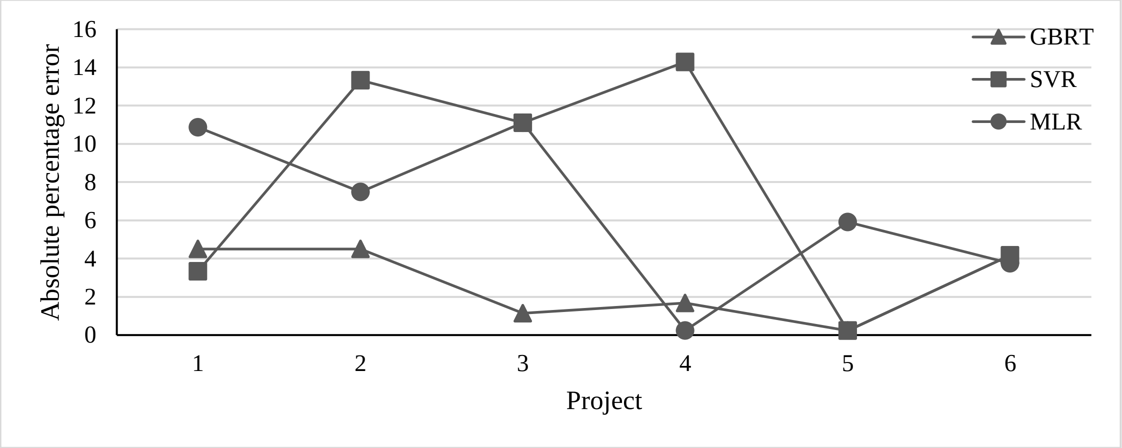


Figure 6.43. Comparison of absolute percentage error (APE) of regularized GBRT with SVR and MLR models for construction schedule prediction (Low rise building).

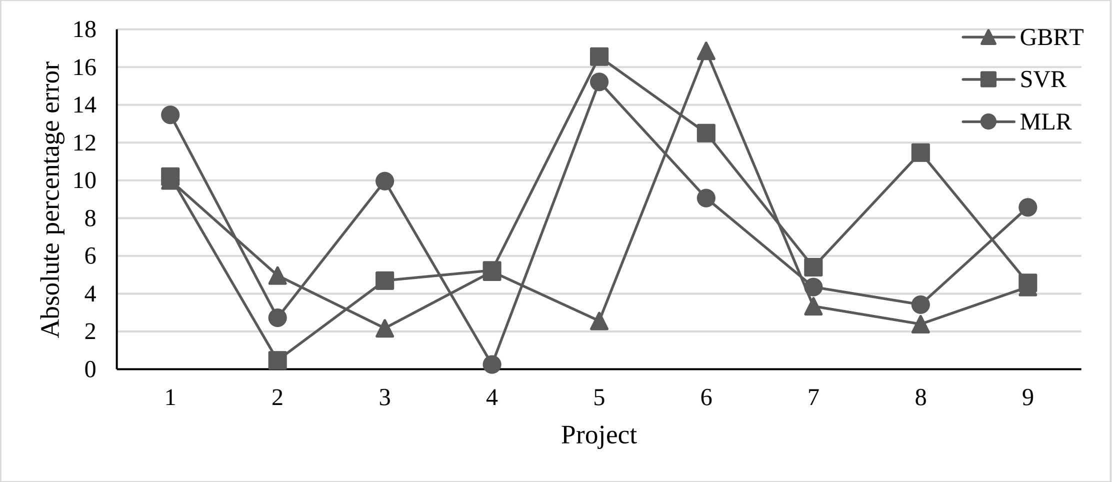


Figure 6.44. Comparison of absolute percentage error (APE) of regularized GBRT with SVR and MLR models for construction schedule prediction (Medium rise building).

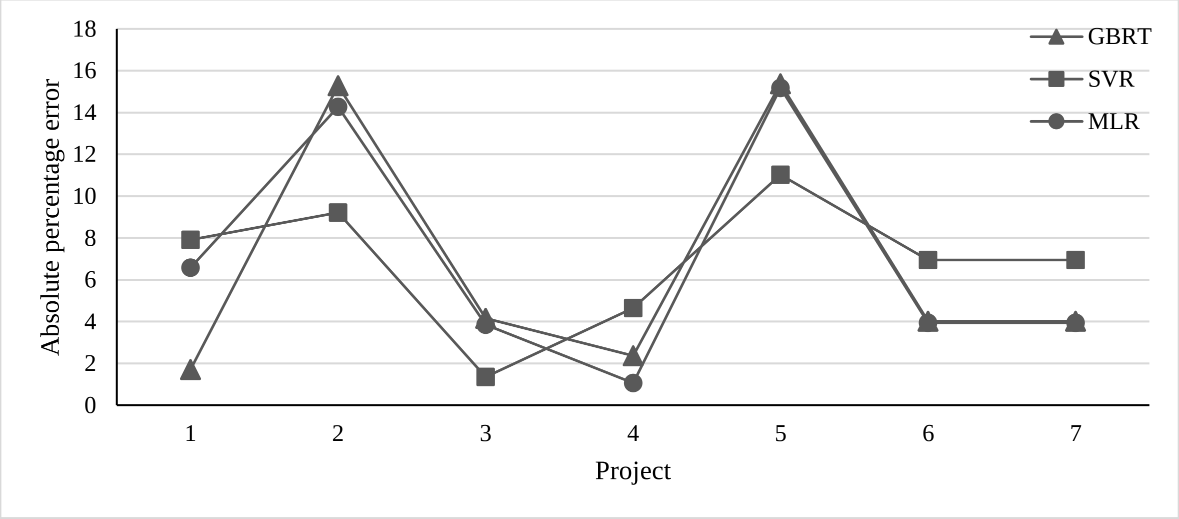


Figure 6.45. Comparison of absolute percentage error (APE) of regularized GBRT with SVR and MLR models for construction schedule prediction (High rise building).

Figure 6.46, Figure 6.47 and Figure 6.48 show the absolute percentage errors (MAPE) for the testing data of low rise, medium rise and high rise buildings for regularized GBRT, Support Vector Regression and Multiple Linear Regression cost prediction models.

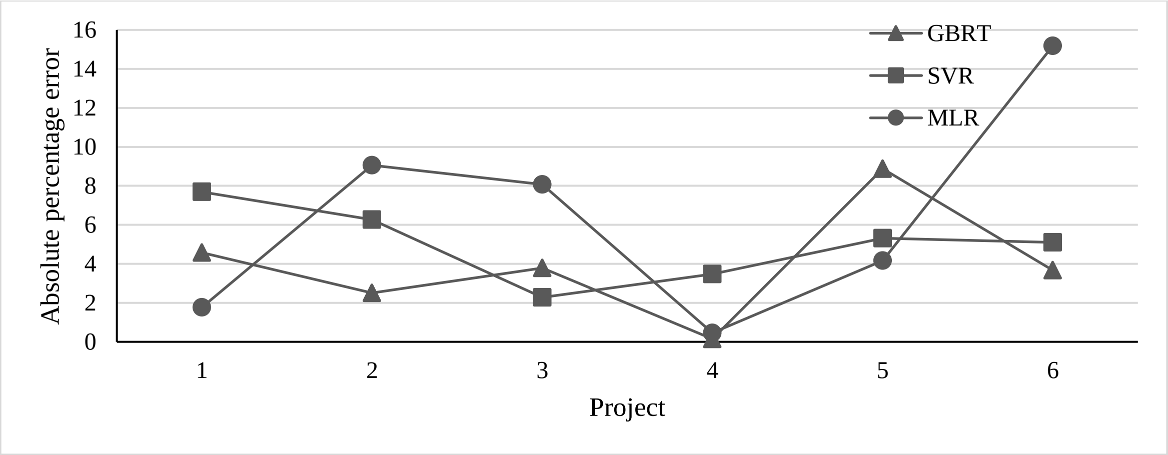


Figure 6.46. Comparison of absolute percentage error (APE) of regularized GBRT with SVR and MLR models for construction cost prediction (Low rise building).

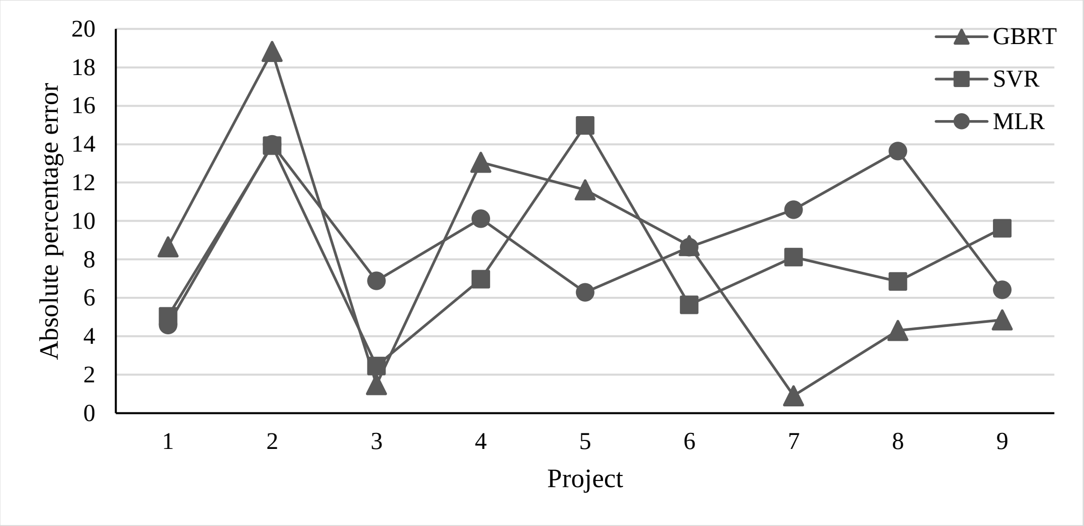


Figure 6.47. Comparison of absolute percentage error (APE) of regularized GBRT with SVR and MLR models for construction cost prediction (Medium rise building).

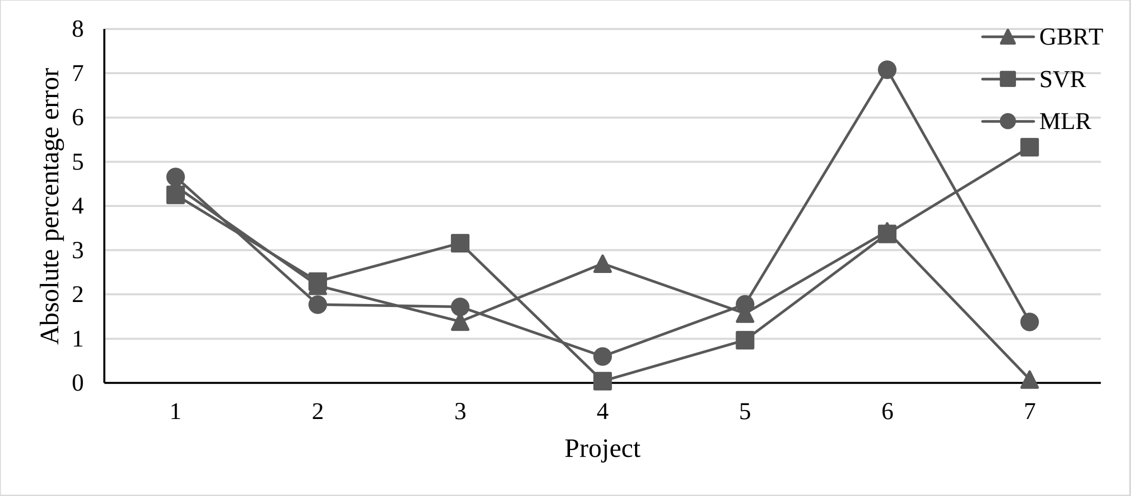


Figure 6.48. Comparison of absolute percentage error (MAPE) of regularized GBRT with SVR and MLR models for construction cost prediction (High rise building).

Table 6.1 shows the summary of performance comparison of regularized GBRT with SVR and MLR models for schedule prediction. It is noted that the average MAPEs for regularized GBRT schedule prediction model are 2.71% for low rise, 5.76% for medium rise and 6.71% for high rise buildings. On the other hand, the MAPEs for Support Vector Regression schedule prediction models of low, medium and high rise buildings are 7.75%, 7.90% and 6.86% respectively. Again, the MAPEs for Multiple Linear Regression schedule prediction models of low, medium and high rise buildings are 6.57%, 7.45% and 6.97% respectively. This indicates that regularized GBRT has performed better than Support Vector Regression and Multiple Linear Regression in schedule prediction. Since, GBRT is a combination of many models it has given better performances than SVR and MLR. From the table, it is seen that the mean squared errors (MSEs) of regularized GBRT models are smaller than MSEs of SVR and MLR models for low rise, medium rise and high rise buildings.

Table 6.1: Summary of performance comparison of schedule prediction models.

	Low rise building		Medium rise building		High rise building	
	MAPE	MSE	MAPE	MSE	MAPE	MSE
Regularized GBRT	2.71%	0.9715	5.76%	20.0077	6.71%	23.2011
Support vector regression (SVR)	7.75%	3.1612	7.90%	23.1612	6.86%	24.4520
Multiple linear regression (MLR)	6.57%	6.0487	7.45%	22.0487	6.97%	25.8164

Table 6.2 represents the summary of performance comparison of regularized GBRT with SVR and MLR models for cost prediction. It is noted that the average MAPEs for regularized GBRT cost prediction models are 3.93% for low rise, 8.05% for medium rise and 2.26% for high rise buildings. On the other hand, the MAPEs for Support Vector Regression cost prediction models of low, medium and high rise buildings are 5.02%, 8.18% and 2.77% respectively. Again, the MAPEs for Multiple Linear Regression cost prediction models of low, medium and high rise buildings are 6.45%, 9.02% and 2.71% respectively. This indicates that regularized GBRT has performed better than Support Vector Regression and Multiple Linear Regression in cost prediction. Here, also the mean squared errors (MSEs) of regularized GBRT cost prediction models are smaller than MSEs of SVR and MLR models for low rise, medium rise and high rise buildings.

Table 6.2: Summary of performance comparison of cost prediction models.

	Low rise building		Medium rise building		High rise building	
	MAPE	MSE	MAPE	MSE	MAPE	MSE
Regularized GBRT	3.93%	0.5507	8.05%	2.9834	2.26%	0.3034
Support vector regression (SVR)	5.02%	1.1832	8.18%	2.2180	2.77%	0.4587
Multiple linear regression (MLR)	6.45%	1.4562	9.02%	3.2354	2.71%	0.5221

CHAPTER 7: CONCLUSIONS AND FUTURE WORK

7.1 Conclusions

The aim of the current research was to develop regularized Gradient Boosted Regression Tree models to predict the construction schedule and cost for low rise, medium rise and high rise buildings. In this thesis, the theory of Gradient Boosted Regression Tree (GBRT) has been described and the regularization of the hyper-parameters of GBRT has been performed by using Random Search method. Regularized GBRT has been applied to estimate the construction schedule and construction cost. The following conclusions can be drawn from the research work.

i. For low rise building, the most significant features, selected through one-way ANOVA F-test, for construction schedule and cost prediction models are “Location (F1)”, “Land size (F2)”, “Floor height (F3)”, “Floor area (F4)”, “No. of basement (F5)”, “Workforce level (F9)”, and “No. of floor (F10)”.

ii. For medium rise and low rise buildings, the most significant features are “Land size (F2)”, “Floor area (F4)”, “No. of basement (F5)”, “Workforce level (F9)”, and “No. of floor (F10)”

iii. Random Search method has been applied to identify the best values of the hyper-parameters to regularize the GBRT models for low rise, medium rise and high rise buildings. In this study, number of trees, minimum sample leaf, maximum depth and learning rate have been tuned for both construction duration and cost prediction model.

iv. For GBRT schedule prediction model of low rise building, the number of trees are 100, minimum sample leaf is 5, maximum depth is 3 and learning rate is 0.6. For GBRT cost prediction model of low rise building, the number of trees are 75, minimum sample leaf is 4, maximum depth is 6 and learning rate is 0.4.

v. For GBRT schedule prediction model of medium rise building, the number of trees are 100, minimum sample leaf is 5, maximum depth is 7 and learning rate is 0.1. For GBRT cost prediction model of medium rise building, number of trees are 125, minimum sample leaf 5, maximum depth is 5 and learning rate is 0.4.

vi. For GBRT schedule prediction model of high rise building, the number of trees are 100, minimum sample leaf is 4, maximum depth is 6 and learning rate is 0.6. For GBRT cost prediction model, the number of trees are 125, minimum sample leaf is 6, maximum depth is 5 and learning rate is 0.5.

vii. MAPEs of regularized GBRT schedule prediction models for low rise, medium rise and high rise buildings are 2.71%, 5.76% and 6.71%. MSEs of the models are 0.9715, 20.0077 and 23.2011. Regularized GBRT model has outperformed SVR and MLR models in schedule prediction

viii. MAPEs of regularized GBRT cost prediction models for low rise, medium rise and high rise buildings are 3.93 %, 8.05% and 2.26%. MSEs of the models are 0.5507, 2.9834 and 0.3034. Regularized GBRT model has outperformed SVR and MLR models in cost prediction

7.2 Future Work

i. In this study, regularized GBRT based robust models for construction schedule and cost prediction have been developed for 69 projects. In future, more projects can be considered to develop more robust model.

ii. For regularization, Random Search method has been used. Evolutionary algorithms like genetic algorithm, particle swarm algorithm (PSO) can be used to optimize the hyper-parameter value in future work.

REFERENCES

Abu Hammad, A. A., Alhaj Ali, S. M., Sweis, G. J., & Bashir, A. (2008). Prediction Model for Construction Cost and Duration in Jordan. In *Jordan Journal of Civil Engineering*, *2*(3), 250-266.

Agrawal, R. K., Muchahary, F., & Tripathi, M. M. (2019). Ensemble of relevance vector machines and boosted trees for electricity price forecasting. *Applied Energy*, *250*, 540–548.

Ahn, J., Park, M., Lee, H. S., Ahn, S. J., Ji, S. H., Song, K., & Son, B. S. (2017). Covariance effect analysis of similarity measurement methods for early construction cost estimation using case-based reasoning. *Automation in Construction*, *81*, 254–266.

Al-Momani, A. H. (1996). Construction cost prediction for public school buildings in Jordan. *Construction Management and Economics*, *14*(4), 311–317.

Amit, Y., & Geman, D. (1997). Shape Quantization and Recognition with Randomized Trees. *Neural Computation*, *9*(7), 1545–1588.

Arage, S. S., & Dharwadkar, N. V. (2017). Cost estimation of civil construction projects using machine learning paradigm. *Proceedings of the International Conference on IoT in Social, Mobile, Analytics and Cloud, I-SMAC 2017*, 594–599.

Bala, K., Bustani, S. A., & Waziri, B. S. (2014). A computer-based cost prediction model for institutional building projects in Nigeria an artificial neural network approach. *Journal of Engineering, Design and Technology*, *12*(4), 518–529.

Batselier, J., & Vanhoucke, M. (2017). Improving project forecast accuracy by integrating earned value management with exponential smoothing and reference class forecasting. *International Journal of Project Management*, *35*(1), 28–43.

Bayram, S. (2017). Duration prediction models for construction projects: In terms of cost or physical characteristics? *KSCE Journal of Civil Engineering*, *21*(6), 2049–2060.

Bayram, S., Ocal, M. E., Laptali Oral, E., & Atis, C. D. (2016). Comparison of multi layer perceptron (MLP) and radial basis function (RBF) for construction cost estimation: the case of Turkey. *Journal of Civil Engineering and Management*, *22*(4), 480–490.

Belavagi, M. C., & Muniyal, B. (2016). Performance Evaluation of Supervised Machine

Learning Algorithms for Intrusion Detection. *Procedia Computer Science*, *89*, 117–123.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*, 281–305.

Boon, L. H., Saar, C. C., Lau, S. E. N., Aminudin, E., Zakaria, R., Hamid, A. R. A., Sarbini, N. N., & Zin, R. M. (2019). Building information modelling integrated project delivery system in Malaysia. *Malaysian Construction Research Journal*, *6*(Special issue 1), 144–152.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Chen, H. L. (2018). Early Prediction of Project Duration: A Longitudinal Study. *EMJ - Engineering Management Journal*, *30*(3), 191–202.

Cheng, M. Y., & Hoang, N. D. (2014). Interval estimation of construction cost at completion using least squares support vector machine. *Journal of Civil Engineering and Management*, *20*(2), 223–236.

Cheng, M. Y., Peng, H. S., Wu, Y. W., & Chen, T. L. (2010). Estimate at completion for construction projects using evolutionary support vector machine inference model. *Automation in Construction*, *19*(5), 619–629.

Choi, H., Son, H., & Kim, C. (2018). Predicting financial distress of contractors in the construction industry using ensemble learning. *Expert Systems with Applications*, *110*, 1–10.

Czarnigowska, A., & Sobotka, A. (2013). Time-cost relationship for predicting construction duration. *Archives of Civil and Mechanical Engineering*, *13*(4), 518–526.

El-Dash, K., Ramadan, O., & Youssef, W. . (2019). Duration Prediction Models for Construction Projects in Middle East. *Engineering, Technology & Applied Science Research*, *9*(2), 3924–3932.

Elfahham, Y. (2019). Estimation and prediction of construction cost index using neural networks, time series, and regression. *Alexandria Engineering Journal*, *58*(2), 499–506.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis*, *38*(4), 367–378.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *Support Vector Machines and Flexible*

Discriminants, 1–42.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(8), 832–844.

Hu, Y., Feng, B., Mo, X., Zhang, X., Ngai, E. W. T., Fan, M., & Liu, M. (2015). Cost-sensitive and ensemble-based prediction model for outsourced software project risk prediction. *Decision Support Systems*, *72*, 11–23.

Jin, Runzhi, Cho, K., Hyun, C., & Son, M. (2012). MRA-based revised CBR model for cost prediction in the early stage of construction projects. *Expert Systems with Applications*, *39*(5), 5214–5222.

Jin, RunZhi, Han, S., Hyun, C., & Kim, J. (2014). Improving Accuracy of Early Stage Cost Estimation by Revising Categorical Variables in a Case-Based Reasoning Model. *Journal of Construction Engineering and Management*, *140*(7), 04014025.

Juszczyk, M. (2019). On the Search of Models for Early Cost Estimates of Bridges: An SVM-Based Approach. *Buildings*, *10*(1), 2.

Juszczyk, M., & Leśniak, A. (2019). Modelling Construction Site Cost Index Based on Neural Network Ensembles. *Symmetry*, *11*(3), 411.

Khamooshi, H., & Abdi, A. (2017). Project Duration Forecasting Using Earned Duration Management with Exponential Smoothing Techniques. *Journal of Management in Engineering*, *33*(1), 04016032.

Kim, G.-H., Shin, J.-M., Kim, S., & Shin, Y. (2013). Comparison of School Building Construction Costs Estimation Methods Using Regression Analysis, Neural Network, and Support Vector Machine. *Journal of Building Construction and Planning Research*, *01*(01), 1–7.

Kim, G. H., An, S. H., & Kang, K. I. (2004). Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. *Building and Environment*, *39*(10), 1235–1242.

Kim, Y.-J., Yeom, D.-J., & Kim, Y. S. (2019). Development of construction duration prediction model for project planning phase of mixed-use buildings. *Journal of Asian Architecture and Building Engineering*, *18*(6), 586–598.

Kohli, P. P. S., Zargar, S., Arora, S., & Gupta, P. (2019). Stock prediction using machine learning algorithms. In *Advances in Intelligent Systems and Computing*, *698*, 405–414.

Koo, ChoongWan, Hong, T., Hyun, C., & Koo, K. (2010). A CBR-based hybrid model for predicting a construction duration and cost based on project characteristics in multi-family housing projects. *Canadian Journal of Civil Engineering*, *37*(5), 739–752.

Koo, Choongwan, Hong, T., Jeong, K., & Kim, J. (2018). Development of the monthly average daily solar radiation map using A-CBR, FEM, and kriging method. *Technological and Economic Development of Economy*, *24*(2), 489–512.

Kumar, M., Rath, N. K., Swain, A., & Rath, S. K. (2015). Feature Selection and Classification of Microarray Data using MapReduce based ANOVA and K-Nearest Neighbor. *Procedia Computer Science*, *54,* 301–310.

Leu, S. Sen, & Liu, C. M. (2016). Using principal component analysis with a back-propagation neural network to predict industrial building construction duration. *Journal of Marine Science and Technology (Taiwan)*, *24*(2), 82–90.

Lhee, S. C., Flood, I., & Issa, R. R. A. (2014). Development of a two-step neural network-based model to predict construction cost contingency. *Journal of Information Technology in Construction*, *19*(September), 399–411.

Lin, T., Yi, T., Zhang, C., & Liu, J. (2019). Intelligent prediction of the construction cost of substation projects using support vector machine optimized by particle swarm optimization. *Mathematical Problems in Engineering*, *2019,* 1-10.

Lowe, D. J., Emsley, M. W., & Harding, A. (2006). Predicting Construction Cost Using Multiple Regression Techniques. *Journal of Construction Engineering and Management*, *132*(7), 750–758.

Luu, V. T., & Kim, S.-Y. (2009). Neural network model for construction cost prediction of apartment projects in Vietnam. *Korean Journal of Construction Engineering and Management*, *10*(3), 139–147.

Mackova, D., Kozlovska, M., Baskova, R., Spisakova, M., & Krajnikova, K. (2017). Construction-duration prediction model for residential buildings in Slovak republic based on computer simulation. *International Journal of Applied Engineering Research*, *12*(13), 3590–3599.

Magdum, S. K., & Adamuthe, A. C. (2017). Construction Cost Prediction Using Neural Networks. *ICTACT Journal on Soft Computing*, *8*(1), 1549–1556.

Mahamid, I. (2019). *The Development of Regression Models for Preliminary Prediction of Road Construction Duration*. *3*(4), 14–20.

Mao, S., & Xiao, F. (2019). Time Series Forecasting Based on Complex Network Analysis. *IEEE Access*, *7*, 40220–40229.

Mensah, I., Adjei-Kumi, T., & Nani, G. (2016). Duration determination for rural roads using the principal component analysis and artificial neural network. *Engineering, Construction and Architectural Management*, *23*(5), 638–656.

Mohan, A., Chen, Z., & Weinberger, K. (2011). Web-search ranking with initialized gradient boosted regression trees. *Journal of Machine Learning Research*, *14*, 77–89.

Naik, M. G., & Radhika, V. S. B. (2015). Time and Cost Analysis for Highway Road Construction Project Using Artificial Neural Networks. *Journal of Construction Engineering and Project Management*, *5*(1), 26–31.

Ng, S. T., Cheung, S. O., Skitmore, M., & Wong, T. C. Y. (2004). An integrated regression analysis and time series model for construction tender price index forecasting. *Construction Management and Economics*, *22*(5), 483–493.

Niu, D., & Hua, F. (2016). Research on Prediction of Transmission and Transformation Project Cost Index Based on ARIMA and Exponential Smoothing Models. In *Proceedings of the 22nd International Conference on Industrial Engineering and Engineering Management 2015*, 771–779.

Niu, H., Gerstoft, P., & Reeves, E. (2017). Statistical inference for source localization using multi-frequency machine learning. *The Journal of the Acoustical Society of America*, *141*(5), 3590–3590.

Persson, C., Bacher, P., Shiga, T., & Madsen, H. (2017). Multi-site solar power forecasting using gradient boosted regression trees. *Solar Energy*, *150*, 423–436.

Petroutsatou, K., Georgopoulos, E., Lambropoulos, S., & Pantouvakis, J. P. (2012). Early cost estimating of road tunnel construction using neural networks. *Journal of Construction Engineering and Management*, *138*(6), 679–687.

Petrusheva, S., Car-Pušić, D., & Zileska-Pancovska, V. (2019). Support Vector Machine Based Hybrid Model for Prediction of Road Structures Construction Costs. *IOP Conference Series: Earth and Environmental Science*, *222*(1), 012010.

Ponomareva, N., Radpour, S., Hendry, G., Haykal, S., Colthurst, T., Mitrichev, P., & Grushetsky, A. (2017). TF Boosted Trees: A Scalable TensorFlow Based Framework for Gradient Boosting. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10536 LNAI*, 423–427.

Principe, J. C. (2010). Information Theoretic Learning Acknowledgments. *Computer Engineering*, *32*(1), 103-134.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, *5*(2), 197–227.

Shayboun, M., & Koch, C. (2019). Sorting things out? Machine learning in complex construction projects. *Proceedings of the 2019 European Conference on Computing in Construction*, *1*, 65–74.

Shr, J.-F., & Chen, W.-T. (2006). FUNCTIONAL MODEL OF COST AND TIME FOR HIGHWAY CONSTRUCTION PROJECTS. In *Journal of Marine Science and Technology*, *14*(3), 127-138.

Sonmez, R. (2011). Range estimation of construction costs using neural networks with bootstrap prediction intervals. *Expert Systems with Applications*, *38*(8), 9913–9917.

Soto, B. G. De, & Adey, B. T. (2015). Investigation of the Case-based Reasoning Retrieval Process to Estimate Resources in Construction Projects. *Procedia Engineering*, *123*, 169–181.

Strom, H., Albury, S., & Sorensen, L. T. (2019). Machine Learning Performance Metrics and Diagnostic Context in Radiology. *11th CMI International Conference, 2018: Prospects and Challenges Towards Developing a Digital Economy within the EU, PCTDDE 2018*, 56–61.

Sullivan, J., Asmar, M. El, Chalhoub, J., & Obeid, H. (2017). Two Decades of Performance Comparisons for Design-Build, Construction Manager at Risk, and Design-Bid-Build: Quantitative Analysis of the State of Knowledge on Project Cost, Schedule, and Quality. *Journal of Construction Engineering and Management*, *143*(6), 04017009.

Thomas, N., & Thomas, A. V. (2016). Regression Modelling for Prediction of Construction Cost and Duration. *Applied Mechanics and Materials*, *857*, 195–199.

Torres-Barrán, A., Alonso, Á., & Dorronsoro, J. R. (2019). Regression tree ensembles for wind energy and solar radiation prediction. *Neurocomputing*, *326–327*, 151–160.

Ugur, L. O., Kanit, R., Erdal, H., Namli, E., Erdal, H. I., Baykan, U. N., & Erdal, M. (2019). Enhanced Predictive Models for Construction Costs: A Case Study of Turkish Mass Housing Sector. *Computational Economics*, *53*(4), 1403–1419.

Wang, Y. R., Yu, C. Y., & Chan, H. H. (2012). Predicting construction cost and schedule success using artificial neural networks ensemble and support vector machines classification models. *International Journal of Project Management*, *30*(4), 470–478.

Wauters, M., & Vanhoucke, M. (2014). Support Vector Machine Regression for project control forecasting. *Automation in Construction*, *47*, 92–106.

Waziri, B. S., Bala, K., & Bustani, S. A. (2017). Artificial Neural Networks in Construction Engineering and Management. *International Journal of Architecture, Engineering and Construction*, *6*(1), 50–60.

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

Willmott, C., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*(1), 79–82.

Yi, T., Zheng, H., Tian, Y., & Liu, J. P. (2018). Intelligent Prediction of Transmission Line Project Cost Based on Least Squares Support Vector Machine Optimized by Particle Swarm Optimization. *Mathematical Problems in Engineering, 2018,* 1-10.

Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*, *58*(9), 308–324.

Appendix

A. Dataset

Project Name	Location	Land size	Floor area	No. of basements	No. of floor	Work force level	Duration	Cost
Project 1	1	1	1	1	9	1	31	12.5
Project 2	1	1	2	2	8	2	33	16.5
Project 3	1	2	2	3	12	2	38	15.5
Project 4	1	2	2	3	12	3	48	17
Project 5	1	2	3	3	11	4	49	19.5
Project 6	1	2	3	3	11	3	44	17
Project 7	1	2	3	3	13	5	52	19
Project 8	1	2	3	2	12	3	45	16.5
Project 9	2	2	3	2	11	4	39	18
Project 10	2	2	3	2	10	2	37	16.5
Project 11	2	3	3	3	13	4	46	17.5
Project 12	2	3	3	3	15	4	54	19
Project 13	2	3	2	2	13	4	48	19.5
Project 14	2	3	2	2	13	4	45	19
Project 15	2	1	1	1	9	1	28	15
Project 16	2	1	1	1	9	1	31	14.5
Project 17	3	1	1	1	8	1	27	12.5
Project 18	3	1	2	1	9	1	36	13.1
Project 19	3	1	2	2	8	2	31	13.2
Project 20	3	1	1	2	9	2	34	14.5
Project 21	3	1	1	2	8	3	35	13
Project 22	3	1	1	2	8	3	33	12.9
Project 23	3	1	1	2	9	3	34	13.1
Project 24	3	1	1	2	9	4	31	15.5
Project 25	3	2	2	2	13	4	53	19.7
Project 26	3	2	2	2	12	4	45	19.5
Project 27	4	2	2	2	12	2	51	17
Project 28	4	2	3	2	12	3	46	18.5
Project 29	4	2	3	2	11	3	40	17.5
Project 30	4	2	3	3	12	4	46	19
Project 31	4	2	3	3	10	2	40	16.5
Project 32	4	2	3	3	10	3	37	18.2
Project 33	4	2	3	3	13	5	48	21
Project 34	4	2	3	3	12	4	42	17.7
Project 35	4	3	2	2	15	4	50	20.2
Project 36	4	3	2	3	15	4	55	19
Project 37	5	3	2	3	14	5	58	21.2
Project 38	5	3	2	3	16	5	54	21.8
Project 39	5	3	2	3	14	4	49	19.8

Project 40	5	3	2	3	15	5	55	23
Project 41	5	3	2	3	13	4	47	21.3
Project 42	5	3	3	2	14	4	50	19.2
Project 43	5	3	3	2	13	5	55	20
Project 44	5	3	3	3	14	5	51	22.5
Project 45	5	3	3	3	14	4	53	19.5
Project 46	5	3	3	3	13	5	54	20.9
Project 47	5	3	3	3	15	4	60	23.4
Project 48	5	1	1	1	8	1	26	13
Project 49	5	1	1	1	9	1	33	12.8
Project 50	1	1	1	2	8	1	29	12.8
Project 51	1	1	1	2	8	1	30	13
Project 52	1	1	2	2	9	1	31	13.5
Project 53	1	2	2	2	11	2	41	15.5
Project 54	1	2	3	2	11	3	42	18
Project 55	1	2	2	3	12	3	41	18
Project 56	1	2	3	2	12	4	42	19.2
Project 57	1	3	2	3	14	5	52	21.2
Project 58	2	3	2	2	15	3	60	21.8
Project 59	2	3	2	2	15	4	52	20.8
Project 60	2	3	2	3	13	5	50	21.6
Project 61	2	3	2	3	14	5	58	21.8
Project 62	2	3	3	2	15	5	60	22.5
Project 63	2	3	3	2	13	5	51	22
Project 64	2	3	3	2	14	5	55	21.5
Project 65	2	3	3	2	15	5	60	21.8
Project 66	3	3	3	2	13	5	51	22
Project 67	1	3	3	3	15	5	54	23.5
Project 68	3	3	3	3	16	5	57	21.6
Project 69	1	3	3	3	16	5	56	21