Chapter Five: Conclusions and Suggestions for Future Work

The findings and conclusions of this work are presented in this chapter. Additionally, it offers recommendations for the work that should be done in the future.

References

1.5 Aim of Thesis

1. Design a modern method based on machine learning techniques for estimating the cost of road construction in the initial phase of the project with high accuracy by utilizing a data set from real projects that were gathered from the Diyala Governorate of the State of Iraq.
2. Improving the forecasting of road construction costs, by designing a forecasting model that uses the data contained in the other variables. The model's predictions are accurate enough to be used in the real world, easy to implement, and require little training.

1.6 Outline of Thesis

In addition to this chapter, the following chapters are included in the remainder of this thesis's components:

Chapter Two: Theoretical Background

The theoretical resources and methodological approaches that were utilized in the creation of this thesis are discussed in this chapter.

Chapter Three: The Proposed Model

This chapter provides an overview of the many stages involved in the proposed prediction system, including its design and its actualization.

Chapter Four: Experimental Results and Evaluation

The outcomes of the proposed model's implementation, analysis, and testing are presented in this chapter, and the chapter also evaluates these results.

1.3 Conceptual Cost Estimating

The process of developing preliminary cost estimates is an integral part of every project's planning and feasibility investigation. Road construction planning decisions made early on are crucial since they can have the most impact on the project's overall construction cost. The term "conceptual cost estimate" refers to the method of estimating a project's total budget using just high-level, initial ideas about the undertaking. Since many of the variables that will end up influencing the final price tag of a project are unknown at the start, conceptual cost estimates can be difficult [4][15].

1.4 Problem Statement

1. An important step in creating a conceptual cost model is identifying the inputs that will be used. However, the performance of the proposed model can be negatively affected by improper input collection. Therefore, decision-makers can benefit from expert advice when calculating the initial price of roads.
2. The absence of a database for government department projects hindered the evolution of cost estimation models. The lack of an integrated database on previously completed construction projects and the non-use of intelligent tools when estimating the costs of construction projects are two of the most significant barriers and challenges faced by estimators in the Republic of Iraq, this may lead to manipulation when estimating the cost of roads, as well as the possibility of errors when estimating costs for many projects in a short time.
3. The traditional forecasting approaches fall short of expectations when dealing with highly volatile data.

The collection included 4811 samples collected from 300 separate sections of a "three-kilometer road built in Iran's Hyrcanian Forests". The model shows IBL model (R = 0.998, RMSE = 1.4%) SVM model (R = 0.993, RMSE = 2.44%).

Table (1.1) Previous Studies

Ref	Cost Estimation Project	Dataset Location	Dataset Prosperities	ML model	Feature Selection	Deep Learning
Peško 2017	Roads	Serbia	2005-2012 (166)	(ANNs) (SVM)	×	×
AL-Zwainy 2017	highway	Iraq	×	ANN	×	×
Rafiei 2018	Building	Iran	1993-2005 (372)	SVM and BPNN	×	DBM
Ogungbile 2018	Road	Nigeria	2010–2015 (151)	linear and multiple regression	×	×
Cao 2018	highway	Georgia	2008 to2016 (1400)	Gbm Xgb Rngr ANN	Boruta feature analysis	×
Barros 2018	highway	Brazil	2010 to 2016 (14)	ANN	×	×
Hakami 2019	Building	Yemen	2011 – 2015 (136)	ANN	×	×
Tijanić 2020	Roads	Croatia	1999 - 2019	ANN	×	×
Mahdavian 2021	Roads	U.S	2001 – 2017 (14,076)	DT, RF, KNN ANN LR,	(RF) , (BR) (DT)	×
Mahalakshmi 2021	Roads	Iran	4811	(LR), K-Star, (MLP), (SVM), (IBL)	×	×

one had the following characteristics: Complexity, The Nature of the Project, Floor space, Details such as the building's height, material of construction, and number of elevators are all important. Surfaced in a slab-like fashion, Form of outer covering Home adornment, A/C Unit Construction Different kinds of tiling, different electrical systems, Definition of a Mechanical Work, below ground level, Ground level, landing zone, Website of the Project. The accurate of the model was MAPE = 0.14 % and R = 0.999.

8. A model of ANN (MLP, GRNN, RBFNN) has been proposed by Tijanić 2020 [12] for Cost estimation in road construction in Croatia, the GRNN has acquired the most reasonable accuracy with MAPE of 0.13 and R^2 of 95%. During the last 20 years, only 57 segments of National roads and highways in the Republic of Croatia were created, and the features used for each were the project's scope and type, the road's length and breadth, the contracted construction duration, and the actual construction expenses.

9. Mahdavian 2021 [13] The forecasting of costs has been automated thanks to a modeling pipeline created by the authors. The dataset for the critical highway construction cost items of the Florida Department of Transportation (FDOT) between 2001 and 2017 has been subjected to feature selection and Machine Learning techniques, comprising 69 different factors such as the housing market, the energy market, social and economic factors, the state of the economy in the United States, and the passage of time. With a 92.51% prediction accuracy, when compared to nonlinear models, the proposed linear model performed exceptionally well in both generalization and prediction of cost components.

10. Mahalakshmi 2021[14] the researchers have developed models to estimate building costs using actual data using machine learning methods.

wearing, the grade of concrete utilized, and the age of the project. Dimensions of haulage, cut-to-spoil depth, and sub-basement depth. “The coefficient of determination R^2 for the developed models ranged from 0.85 to 0.99”.

5. Cao 2018 [9] authors proposed a robust ensemble learning model was “gradient boosting (gbm), extreme gradient boosting (xgb), and random forest for the first level, and ANN was the second level to predict the value of unit price bids of Highway Projects in Georgia”, More than 1,400 projects' worth of bidding information was used in every offer, 57 features have been collected by the Georgia Tech ESBE lab. Information about the local highway construction market, the state of the construction industry, the macroeconomy, and the oil market, as well as the specifics of the project itself, such as its location and its proximity to major suppliers of necessary materials. The ensemble learning model showed higher result than (gbm), (xgb), and random forest with MAE 37.98%.
6. Barros 2018 [10] Using Artificial Neural Networks, the researcher developed a more precise method of estimating highway development projects in Brazil. There were a total of fourteen projects utilized in the training and validation phases, with one used in the testing phase. Mainstream time to execution, metallic material's typical transit distance cement transport distance on average, petroleum asphalt cement transport distance on average, excavation volume, embankment volume, asphalt concrete volume construction totals for bridges, on average, the estimated costs to extend the bridges were 99% accurate.
7. Hakami 2019 [11] used a cutting-edge methodology called an artificial neural network to demonstrate the superiority of this approach over the conventional one. The dataset consisted of 136 finished projects in State Yemen, and each

(construction work and-or reconstruction) Estimation of cost for works on roadway construction and landscaping was low due to the input parameters used to develop the model where MAPE showed result as 7.06 .

2. AL-Zwainy 2017[6] they have use Multi-layer perceptron trainings utilized back-propagation algorithm for predict construction costs highways in Iraq, the dataset was small about 150 projects. In the input model, there were two types of variables: objective variables (length of the highway in (km), capacity, number of interchanges, estimate year, length of major bridges, stream crossing) , and subjective variables (class, material, technology, furnishing, and drainage). Predicting the price of highway project structure works is a breeze with ANNs with degree MAPE = 6.81%, RMSE = 0.30772 and (R^2) was 81.05%.

3. Rafiei 2018 [7]developed a machine learning-based construction cost estimating model that factors in economic variables and indices (EV&Is). The model included a deep Boltzmann machine, backpropagation "neural network BPNN, and support vector machine (SVM)". The 372 condos in Tehran, Iran, ranging in height from 3- to 9-stories, were used to test the proposed paradigm. The buildings were constructed between 1993 and 2008. A typical training accuracy of 95.1–100% for Network 1 (1000 iteration) and the Network 2 for (100 iteration) 87.6% to 90.3%, MSE = 0.043.

4. Ogungbile 2018 [8] constructed cost models employing linear and multivariate regression for forecasting road project estimates. The dataset was (97 project) between 2010 and 2015, 20 separate road construction projects were finished in South-Western Nigeria. When deciding which roads to study, researchers looked at a number of factors, including the amount of asphaltic concrete used in the binder, the amount of asphaltic concrete used in the

Cost models help estimate road construction costs conceptually. Several factors impact project costs, making cost model creation difficult. Prediction process components include project specifics, previous data, current data, estimating approach, cost estimator, and estimates. The inputs to the cost model can only be as accurate as the details provided by the project manager, and here is where the project information comes in. In order to statistically create a cost model, "historical data" are the gathered facts about completed projects in the past. Information on the project has been mined for current data, such as unit labor and material costs and equipment utilization rates. An example of an estimating methodology is the parametric cost model. The cost estimate is derived from the variables to be input or data entered by the "cost estimator", who is the employee of the cost model. These numbers are what the cost model produces as its estimations [4].

1.2 Related Works

Over the years, a number of researchers have undertaken research into construction models with the goal of predicting the preliminary estimate of road projects; this line of inquiry is being driven by the requirement for precise preliminary estimates. The current cost estimation procedures for road projects, particularly those widely used during the conceptual stage, have been described. A variety of research employing machine learning approaches to estimate construction costs have been published:

1. Peško 2017[5] applied AI to the task of cost and time estimation on building projects for greater accuracy. “Two types of Artificial Neural Networks (ANN) models were used, Multi-layer Perceptron (MLP) and Radial Basis Function (RBF) models Classification” issues are dealt with by both models, whereas regression issues are handled by MLP models and clustering issues are handled by RBF models. Dataset was 166 projects for Urban Roads

Chapter One

Introduction

1.1 Overview

The ability to accurately foresee conceptual expenses is a crucial factor in early-stage decision making for civil engineering projects. The governmental and urban construction market is sizable, unstable, and financing-intensive. A significant portion of the construction industry in developing economies is national or municipal road construction. In such economies, infrastructure expansion, road restoration, and construction projects account for a sizeable portion of the national appropriations for international donors' assistance. The community of donors will be encouraged to continue funding the developmental programs if projects are completed on schedule and without additional expense [1].

Therefore, in order to anticipate the necessary cost of projects, it is crucial to describe aspects and factors that significantly affect a project's budget. Time constraints necessitate prompt completion of the forecast. This means that decision-makers, project managers, and cost engineers face a hurdle when trying to accurately estimate costs associated with concepts [2].The estimated price tag of highway, road, and bridge construction is estimated in several different ways. Initially, engineers or employers manually tally up anticipated or predicted project costs. As the world's population has grown, businesses have had to estimate expenses using increasingly inadequate tools like spreadsheets, databases, and static computer programs. These algorithms can't forecast cost. The static technique has a high risk of positive or negative mistake and is time-consuming. Self-learning is a dynamic problem-solving strategy that can help with difficult issues [3].

Chapter One

Introduction

List of Tables

Algorithms Used

List of Figures

List of Abbreviations

Abbreviations	*Full Form*
AI	Artificial Intelligence
ANN	Artificial neural networks
BOQ	Bill of Quantities
BPNN	Backpropagation Neural Network
EDA	Exploratory Data Analysis
GBM	Gradient Boosting
GDP	Gross Domestic Product
GRNN	General Regression Neural Network
IQR	The Interquartile Range
k-NN	k Nearest Neighbors
MAPE	Mean Absolute Percentage Error
MLP	Multi-layer Perceptron
MSE	Mean Square Error
Q-Q plot	Quantile-Quantile Plot
$\mathbf{R^2}$	R-squared
RBF	Radial Basis Function
RBFNN	Radial Basis Function Neural Networks
RMSE	Root Mean Square Error
SCEA	Society of Cost Estimating and Analysis
SPSS	Statistical Product and Service Solutions
SVM	Support Vector Machine
XGB	Extreme Gradient Boosting
Σ	Standard Deviation
μ	Mean

List of Content

A credible road construction dataset containing 1658 projects was chosen. Based on the error scores, the various outcomes of the machine learning algorithms are compared. During the training phase, the Bayesian optimization method was utilized to determine the hyperparameters for the algorithms. The validity of the findings obtained during the training phase was checked with 30 K-fold cross validation tests. The Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and R-squared (R^2) coefficient of determination were utilized to assess the models' performance.

Analysis of the findings of the four proposed models revealed that their performance is excellent and their results are extremely close. Random Forest Regression (RF) is the best model with results R^2 = 0.81, MAPE =0.007, and RMSE=0.014.

The machine learning results demonstrated the significant impact of hyperparameters tuning on the performance of the proposed models, where the value of RMSE = 0.0003 for the K-nearest regressor (KNR) and the Ridge regression models RMSE = 0.0004. The overall findings revealed that the estimations provided here were correct and in accordance with the project proposals. As a result, these models might be used as a guideline for allocating financial resources effectively in the early stages of the bidding process.

Abstract

The design of an accurate model of parametric cost during the project conceptual phase represents a critical issue faced by any project manager and decision-maker. Many existing statistical and probabilistic algorithms have been developed to be utilized for predicting projects' costs. However, these developed algorithms can provide inaccurate results owing to utilizing unstable and small samples of data. Recently, various effective models based on Artificial intelligence (AI) techniques have emerged for the applications of supervised regression.

Case databases are few, and many of the proposed ones were ineffective. As a result, this research recommends a new data set that incorporates road construction features and economic benefits in the year of project development. Using machine learning (ML) techniques to train an accurate predictive model by using actual project data for roads in the State of Iraq / Diyala Governorate for the years 2012 through 2021.

This study proposed a model for machine learning capable of predicting road construction costs. The proposed model has five phases: First, collect data and create road construction datasets. Second, these datasets are analyzed using Exploratory Data Analysis (EDA) to identify duplicate rows, missing values, and outliers. Features preprocessing is the third stage. In the fourth stage, the model is trained using four distinct algorithms and then evaluated.

The most important methods include (EDA), which is used to identify and eliminate outlier data, and Pearson Correlation Coefficient, which is applied to determine the important characteristics by employing a correlation objective that is more than (0.12).

Acknowledgments

Praise is to Allah the lord of the whole creation on all the blessings that helped achieve this research.

I would represent my thanks to my first supervisor Dr. Taha Mohammad Hassan for supervising this research and for his generosity, patience, and continuous guidance throughout the work, it has been my good luck to have advice and guidance from him. Many thanks to my second supervisor Dr. Raquim Nihad Zehawi for his valuable assistance in achieving this research. I would express my thanks to the academic and administrative staff at the Department of the computer sciences\University of Diyala.

I want to express my appreciation to my mother. Without her constant prayers, I would not be here, and to the souls of my father and brother. There are no words sufficient to express my gratitude to my sisters, their daughters, and their sons for supporting me all the time.

I extend my sincere thanks to the Director of Roads and Bridges Department in Diyala Governorate, Eng. Hani Ghazi Fakhri, and Head of Planning and Design Department, Eng. Suzan Muhammad Yaqoub, and Eng. Tahseen Abbas Hussain for their valuable assistance, this research would not have been possible without their valuable help.

Yasamin Gh. Abed

Dedication

To...

My mother

Soul of my father and my brother

My dear sisters and their children

All our distinguished teachers those who paved the way for our science and knowledge

To all My Friends.

I produce this work with all my love…

Yasamin Gh. Abed

بسم الله الرحمن الرحيم

﴿ وَلَقَدْ أَوْحَيْنَا إِلَىٰ مُوسَىٰ أَنْ أَسْرِ بِعِبَادِي فَاضْرِبْ لَهُمْ طَرِيقًا فِي
الْبَحْرِ يَبَسًا لَّا تَخَافُ دَرَكًا وَلَا تَخْشَىٰ ﴾

صدق الله العظيم

سورة طه الاية (77)

Ministry of Higher Education and Scientific Research

University of Diyala

College of Science

Department of Computer Science

Predictive Modeling of Road Construction Costs Using Machine Learning Approach

A thesis

Submitted to the Department of Computer Science \ College of the Science \ University of Diyala in a Partial Fulfillment of the Requirements for the Degree of Master in Computer Science

By

Yasamin Ghadbhan Abed

B.Sc. Computer Science Dept. /Diyala University 2009

Supervised By

Prof. Dr . Taha Mohammed Hasan

Prof. Dr. Raquim N. Zehawi

2023 A.D **1444 A.H.**