GEORGIA DOT RESEARCH PROJECT 20-17

Final Report

ENHANCING THE ACCURACY OF CONSTRUCTION COST ESTIMATES FOR MAJOR LUMP SUM (LS) PAY ITEMS AND GENERATING A MORE-ACCURATE LIST OF PAY ITEMS THROUGHOUT THE DESIGN DEVELOPMENT PROCESS

Office of Performance-based Management and Research
600 West Peachtree Street NW | Atlanta, GA 30308

June 2022

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No.
FHWA-GA-22-2017

2. Government Accession No.
N/A

3. Recipient's Catalog No
N/A

4. Title and Subtitle
Enhancing the Accuracy of Construction Cost Estimates for Major Lump Sum (LS) Pay Items and Generating a More-Accurate List of Pay Items Throughout the Design Development Process

5. Report Date
June 2022

6. Performing Organization Code
N/A

7. Author(s)
Baabak Ashuri, Ph.D., DBIA
Minsoo Baek, Ph.D.
Mingshu Li

8. Performing Organ. Report No.
20-17

9. Performing Organization Name and Address
Economics of Sustainable Built Environment (ESBE) Lab
Georgia Institute of Technology
280 Ferst Drive, Atlanta, GA 30332-0680

10. Work Unit No.
N/A

11. Contract or Grant No.
PI#0017413

12. Sponsoring Agency Name and Address
Georgia Department of Transportation (SPR),
Office of Performance-based
Management and Research,
600 W. Peachtree Street NW
Atlanta, GA 30308

13. Type of Report and Period Covered:
Final Report (August 2020 – June 2022)

14. Sponsoring Agency Code
N/A

15. Supplementary Notes
Conducted in cooperation with the U.S. Department of Transportation, Federal Highway Administration.

16. Abstract

State departments of transportation (DOTs) encounter a critical challenge in estimating accurate cost estimates for major lump sum (LS) pay items, such as Traffic Control and Grading Complete, due to incomplete project information during the early stages of project development. To estimate prices for LS pay items, cost estimators and designers in state DOTs apply engineering judgment using knowledge from similar projects from the past and reach out to subject matter experts for providing additional resources. However, researching similar projects for finding appropriate estimates for the LS pay item is not a simple endeavor and leads to significant inaccuracy of cost estimates. A need exists to develop new methods that are capable of capturing key information from project documents and incorporating the complex nonlinear relationships between input and output variables in developing prediction models for LS pay item prices for highway projects. Thus, the overarching objective of this research is to develop forecasting models to estimate the prices of the Traffic Control and Grading Complete LS pay items using advanced text mining and machine learning algorithms that detect key patterns of information generated during project development and provide higher accuracy in cost estimates. In this research, a forecasting model for the prices of the Traffic Control and Grading Complete LS pay items was developed using machine learning algorithms (i.e., random forest, bagging, k-nearest neighbors, and stacking regressor). Furthermore, a web-based application tool was developed in a Python environment to help designers developing cost estimates with a data-driven tool for estimating the prices of the Traffic Control and Grading Complete LS pay items.

17. Key Words
Grading Complete, Lump Sum Pay Item, Machine Learning Algorithms, Text Mining, Traffic Control

18. Distribution Statement
No Restrictions

19.Security Classification (of this report)
Unclassified

20. Security Classification (of this page)
Unclassified

21. Number of Pages
82

22. Price
Free

GDOT Research Project 20-17

Final Report

ENHANCING THE ACCURACY OF CONSTRUCTION COST ESTIMATES FOR MAJOR LUMP SUM (LS) PAY ITEMS AND GENERATING A MORE-ACCURATE LIST OF PAY ITEMS THROUGHOUT THE DESIGN DEVELOPMENT PROCESS

By

Baabak Ashuri, Ph.D., DBIA
Professor of Building Construction, and Civil and Environmental Engineering

Minsoo Baek, Ph.D.
Assistant Professor of Construction Management

Mingshu Li
Graduate Student

Georgia Institute of Technology
Colleges of Design and Engineering

Kennesaw State University
College of Architecture and Construction Management

Contract with
Georgia Department of Transportation

In cooperation with
U.S. Department of Transportation
Federal Highway Administration

June 2022

The contents of this report reflect the views of the authors who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the Georgia Department of Transportation or the Federal Highway Administration. This report does not constitute a standard, specification, or regulation.

SI* (MODERN METRIC) CONVERSION FACTORS

APPROXIMATE CONVERSIONS TO SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
		LENGTH		
in	inches	25.4	millimeters	mm
ft	feet	0.305	meters	m
yd	yards	0.914	meters	m
mi	miles	1.61	kilometers	km
		AREA		
in^2	square inches	645.2	square millimeters	mm^2
ft^2	square feet	0.093	square meters	m^2
yd^2	square yard	0.836	square meters	m^2
ac	acres	0.405	hectares	ha
mi^2	square miles	2.59	square kilometers	km^2
		VOLUME		
fl oz	fluid ounces	29.57	milliliters	mL
gal	gallons	3.785	liters	L
ft^3	cubic feet	0.028	cubic meters	m^3
yd^3	cubic yards	0.765	cubic meters	m^3
		NOTE: volumes greater than 1000 L shall be shown in m^3		
		MASS		
oz	ounces	28.35	grams	g
lb	pounds	0.454	kilograms	kg
T	short tons (2000 lb)	0.907	megagrams (or "metric ton")	Mg (or "t")
		TEMPERATURE (exact degrees)		
^oF	Fahrenheit	5 (F-32)/9 or (F-32)/1.8	Celsius	^oC
		ILLUMINATION		
fc	foot-candles	10.76	lux	lx
fl	foot-Lamberts	3.426	candela/m^2	cd/m^2
		FORCE and PRESSURE or STRESS		
lbf	poundforce	4.45	newtons	N
lbf/in^2	poundforce per square inch	6.89	kilopascals	kPa

APPROXIMATE CONVERSIONS FROM SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
		LENGTH		
mm	millimeters	0.039	inches	in
m	meters	3.28	feet	ft
m	meters	1.09	yards	yd
km	kilometers	0.621	miles	mi
		AREA		
mm^2	square millimeters	0.0016	square inches	in^2
m^2	square meters	10.764	square feet	ft^2
m^2	square meters	1.195	square yards	yd^2
ha	hectares	2.47	acres	ac
km^2	square kilometers	0.386	square miles	mi^2
		VOLUME		
mL	milliliters	0.034	fluid ounces	fl oz
L	liters	0.264	gallons	gal
m^3	cubic meters	35.314	cubic feet	ft^3
m^3	cubic meters	1.307	cubic yards	yd^3
		MASS		
g	grams	0.035	ounces	oz
kg	kilograms	2.202	pounds	lb
Mg (or "t")	megagrams (or "metric ton")	1.103	short tons (2000 lb)	T
		TEMPERATURE (exact degrees)		
^oC	Celsius	1.8C+32	Fahrenheit	^oF
		ILLUMINATION		
lx	lux	0.0929	foot-candles	fc
cd/m^2	candela/m^2	0.2919	foot-Lamberts	fl
		FORCE and PRESSURE or STRESS		
N	newtons	0.225	poundforce	lbf
kPa	kilopascals	0.145	poundforce per square inch	lbf/in^2

* SI is the symbol for the International System of Units. Appropriate rounding should be made to comply with Section 4 of ASTM E380. (Revised March 2003)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

The accuracy of cost estimates during project development highly depends on the extent of information available at the time that the estimate is developed. At the earlier stages of project development (e.g., conceptual or preliminary design stages), at which the design is not complete, and quantities are uncertain, state departments of transportation (DOTs) often encounter difficulty in accurately estimating costs for highway projects. Inaccurate estimates lead to critical issues in delivering projects on time and within budget.

One of the main challenges is the ability to develop accurate cost estimates for major lump sum (LS) pay items, such as Traffic Control and Grading Complete. For example, in the Georgia Department of Transportation (GDOT), Traffic Control and Grading Complete LS pay items are shown on the summary of quantities (SOQ) drawings but do not include individual items of work or quantities that constitute the lump sum measurement and payment, which can cause significant inaccuracy of cost estimates for LS pay items. Cost estimators and designers apply engineering judgment, using knowledge from similar projects from the past, and reach out to subject matter experts for additional resources. However, researching similar projects to find appropriate estimates for the LS pay item is not a simple endeavor. Moreover, the unique features of LS pay items add to the complexity of cost estimating for these items. Therefore, this research project aims to utilize advanced machine learning algorithms to develop appropriate cost

estimating models to enhance the accuracy of early cost estimates for major LS pay items using information from project-specific characteristics, location, and design features.

The overarching objective of this research is to develop forecasting models to estimate the prices of the Traffic Control and Grading Complete LS pay items using advanced text mining and machine learning algorithms that: (1) detect key patterns of information generated during project development and (2) provide higher accuracy of the cost estimates.

To achieve these research objectives, this project uses *text mining algorithms*, including term frequency–inverse document frequency (TF–IDF) and principal component analysis (PCA), to capture key patterns of information from unstructured text files (i.e., concept reports, field plan review reports, preconstruction status reports, and construction plan staging notes). The project then uses *data processing algorithms*, including the synthetic minority oversampling technique (SMOTE) and the Boruta feature selection algorithm, and *machine learning algorithms*, including random forest, bootstrap aggregating (bagging), k-nearest neighbors (KNN), stacking regressor, and ordinary least squares (OLS) linear regression, to develop forecasting models for the prices of the Traffic Control and Grading Complete LS pay items. This study collected the prices of the Traffic Control and Grading Complete LS pay items used in highway projects in the state of Georgia. With the collected data, the researchers developed a forecasting model for the prices of the Traffic Control and Grading Complete LS pay items.

This research project used several machine learning algorithms to develop forecasting models for the segments of the collected data and select the best-performing algorithms for predicting the prices of the Traffic Control and Grading Complete LS pay items for each segment. Based on the mean absolute percentage error (MAPE), the researchers found that the best-performing algorithms for predicting the prices of a Traffic Control LS pay item in Segments 1, 2, 3, 4, and 5, respectively, are: k-nearest neighbors (KNN), random forest, KNN, random forest, and stacking regressor. Moreover, the best-performing algorithms in predicting the prices of a Grading Complete LS pay item in Segments 1, 2, and 3, respectively, are: KNN, KNN, and random forest. Next, the accuracy of the forecasting models is compared between partitioned data and data without partitioning. The model comparison results indicate that the developed machine learning models for forecasting the prices of the Traffic Control and Grading Complete LS pay items in the defined segments show a higher level of forecasting accuracy.

Furthermore, a web-based application tool is developed in a Python environment to help designers developing cost estimates- with a data-driven tool for estimating the prices of the Traffic Control and Grading Complete LS pay items. This tool aids designers and cost estimators with a flexible and intelligent platform for early cost estimation of two important LS line items, Traffic Control and Grading Complete.

CHAPTER 1. INTRODUCTION

INTRODUCTION

Accuracy of cost estimates for highway projects is one of the major components in making decisions for programming and budgeting for a project during the early stages of the project development process. Developing an accurate cost estimate for highway projects during the early stages of project development is burdensome for state departments of transportation (DOTs) because of difficulties in describing scope solutions for all issues, evaluating the quality and completeness of early cost estimates, identifying significant areas of variability and uncertainty in project scope and costs, and tracking the cost impact of design development on major cost estimates (Anderson et al. 2007). State DOTs often encounter critical issues in delivering their projects on time and within budget since an inaccurate cost estimate is the leading root cause of scope changes, cost escalations, project cancellations or delays, and the loss of public trust throughout the project development (Paulsen et al. 2008, Baek and Ashuri 2021).

In addition, the accuracy of cost estimates during project development highly depends on the extent of information available at the time the estimate is developed. At the earlier stages of project development (e.g., the conceptual or preliminary design stages), where the design is not complete and quantities are uncertain, cost estimates are developed using parametric cost factors, such as location, traffic management considerations, and utility impacts, in what is known as a top-down cost estimating technique. Thus, to develop cost estimates at the early stages of the plan development process (PDP), cost estimators often rely on historical bid data due to limited information about project design details (Chou et al. 2006, Gardner et al. 2016). As the complete design and precise quantities of work become available, cost estimates are developed by assigning unit rates for each activity and summing each activity cost to estimate a total construction cost of a project, which is known as a bottom-up estimating technique (Kim et al. 2012, Gardner 2015). With a bottom-up estimating technique, cost estimates can be determined by either prices provided by suppliers or through recent bid history. Ultimately, inaccurate cost estimates for highway projects are still a pervasive problem (Trost and Oberlender 2003, Baek et al. 2016).

PROBLEM STATEMENT

The Georgia Department of Transportation (GDOT) Office of Engineering Services oversees developing engineer estimates for the projects led by the Department and acts as the gatekeeper of cost estimates throughout the PDP milestones. The Office of Engineering Services develops the engineer's estimate for the project using a bottom-up unit cost estimating approach based on pay item quantities taken from the final detailed project plans and specifications. The GDOT Office of Roadway Design, the district design office, and the consultant design phase leader (DPL) each play a unique and important role to initialize the construction cost estimate for the project during the early phases of preliminary engineering (PE). These subject matter experts update the programmed construction cost estimate throughout the Plan Development Process (PDP) milestones and at any time there is a 10% or greater cost increase or decrease. As appropriate or when requested, the Office of Engineering Services assists the DPL in preparing and updating cost estimates.

Developing a reasonably accurate cost estimate during the early phases of the design development is a challenging task for the DPL, as detailed design information has not yet been developed. The DPL must rely on personal experience in similar projects to develop an initial cost estimate. Sometimes, the DPL reaches out to subject matter experts in the Office of Engineering Services for advice and recommendations to prepare and update a more accurate estimate for the project.

One of the main challenges the DPL faces is in developing accurate cost estimates for major lump sum (LS) pay items, such as Traffic Control, Clearing & Grubbing, and Grading Complete. These items are shown on the summary of quantities (SOQ) drawings; however, these drawings do not include individual items of work or quantities that constitute the LS measurement and payment. For instance, according to GDOT Policy 2434-1, Method of Payment for Earthwork, “Grading Complete” (Specification Section 210) is a LS pay item; thus, no quantity, even for information only, is shown on the plans. Grading Complete is automatically used if the total earthwork is not greater than 100,000 CY and includes the Clearing & Grubbing LS items of work.

Developing cost estimates for these LS pay items requires that the DPL apply engineering judgment, use knowledge from similar projects from the past, and reach out to subject matter experts for additional resources. However, researching similar projects to find appropriate estimates for the LS pay item is not a simple endeavor. The unique features of LS pay items add to the complexity of cost estimating for these items, and the price patterns for the LS pay items tend to be more uncertain, compared to other pay items that have well-established records of historical prices. This research aims to utilize advanced machine learning algorithms to develop appropriate cost estimating models to enhance the accuracy of early cost estimates for major LS pay items using information from project-specific characteristics, location, and design features.

Thus, this research develops new methods capable of capturing key information from project documents and incorporating the complex nonlinear relationships between input and output variables in developing reliable prediction models for LS pay item prices for a highway project. This report describes the research process to achieve this objective.

REPORT ORGANIZATION

The report is structured into the following chapters:

- **Chapter 2. Literature** Review**:** Through comprehensive content analysis, this chapter studies recent trends in developing engineers' estimates for transportation projects. The main goal of this task is to collect information and data related to the state of the knowledge about developing cost estimates in early phases of the plan development process.

- **Chapter 3. Research Methodology:** This chapter provides an overview of the research methodology. Appropriate machine learning algorithms devised for data mining and forecasting LS pay items are discussed in this chapter.
- **Chapter 4. Data Collection And Processing:** This chapter reviews the collected data and conducts quality assurance on the collected information, to ensure that the right dataset is used to conduct further analysis. Through a consultation with the Offices of Engineering Services and Roadway Design, the collected data are verified for developing forecasting models for prices of the LS pay items. This critical step ensures the input data are of high quality prior to being incorporated into any modeling efforts.
- **Chapter 5. Development Of Forecasting Models For Lump Sum Pay Items:** Forecasting models are developed through machine learning algorithms. A wide range of modeling techniques that can be potentially applicable for forecasting LS pay items is considered. The quality and accuracy of several modeling choices considering the availability of data points are examined. For example, the following steps are conducted to develop machine learning algorithms (i.e., random forest, bootstrap aggregating [bagging], and k-nearest neighbors [KNN]) for forecasting estimates for LS pay items:

  - Split the dataset into training and testing datasets (this research uses the training dataset to develop the prediction model and validates the developed model by applying it to the testing dataset).
  - Apply several machine learning algorithms, such as random forest, bagging, and k-nearest neighbors, to the training dataset to create forecasting models.
  - Estimate the parameters of the machine learning algorithms through trial-and-error analysis to develop a model with an acceptable level of fitness.

  To validate the accuracy of the developed forecasting models, the best machine learning model is determined based on a predictability assessment. The predictability of each machine learning model is assessed based on computing the difference between the predicted values of the LS pay items and the actual estimates for those line items in the testing dataset. This research uses mean absolute percentage error (MAPE) to evaluate the accuracy/predictability of the developed models.

- **Chapter 6. Web-Based Application For Forecasting Prices Of Lump Sum Pay Items:** A web-based application tool is created to automate the developed forecasting models. This chapter provides a detailed description of how to use the web-based application for forecasting models, alongside snapshots of data entry and results publication. The research team's technical/implementation manager outlines all the required steps to develop an executable tool for forecasting. All steps in the developed algorithms are described in detail to facilitate the implementation of the forecasting model for GDOT.

- **Chapter 7. Conclusions:** A summary of the research findings is presented.

CHAPTER 2. LITERATURE REVIEW

BACKGROUND

Several studies have attempted to enhance the accuracy of cost estimates for transportation construction projects. Past studies have been carried out to estimate and forecast construction costs using historical data through quantitative methods, such as regression analysis and machine learning algorithms. For example, many researchers used various types of linear regression models for forecasting cost estimates for transportation projects. Chou et al. (2006) used multivariate linear regression to predict the unit prices of the major work items (e.g., earthwork and landscape, structures, and subgrade treatments and base). They used project-related parameters, such as the number of work items, project length, and project types, in developing predictive models. Mahamid (2011) used linear regression to predict the total cost of road construction projects. The author developed multivariate regression models to estimate costs of major activities (e.g., earthworks and asphalt works) and sum up the costs of construction activities to calculate the total project cost. Another study conducted by Mahamid (2013) developed regression models to forecast the conceptual cost estimates for road construction projects. That study showed that project-related factors, including bid quantities of the major construction activities, road length, and road width, help predict conceptual cost estimates and the developed regression models provide favorable accuracy in the early stages of a project. Blampied (2018) used multiple regression analysis to estimate conceptual construction costs for public highway projects. The author used 39 pedestrian access facility projects on state highways in California and several input variables, such as the number of ramps, audible traffic signals, and project length. The study showed that the developed exponential regression models were reasonably accurate for forecasting conceptual cost estimates for public highway projects.

Ogungbile et al. (2018) employed a multiple linear regression model to predict the cost of road construction projects. Their study used information related to several quantitative factors, such as asphaltic wear course, earthwork, and site clearance, in developing a cost estimating model. A study conducted by Baek and Ashuri (2018) performed geographical regression analysis to estimate the unit price of asphalt line items used in highway projects. Their study found different linear relationships between the unit prices and external factors (e.g., project length, project types, and duration) depending on the geographical location of a project.

In a follow-up study, Baek and Ashuri (2019) developed a random parameter regression model to estimate the unit price of asphalt line items. The authors identified significant factors representing project characteristics, major supply sources, construction market conditions, macroeconomic conditions, and energy market conditions. Li et al. (2021) investigated the construction cost estimation from the temporal perspective and identified leading indicators of deviation between the owner’s estimate and low bids from the construction market and economic conditions. Another study from Li and Ashuri (2021) quantified the likelihood of construction cost underestimation through Cox proportional hazards regression. However, one of the primary limitations of linear regression analysis is that a defined mathematical form for the cost function is required for better fitting the available historical cost data with explanatory variables (Creese and Li 1995). In addition, it is difficult to account for a large number of variables presented in a construction project and explain the numerous interactions/relationships among variables, which may cause the low accuracy of cost estimating models (de la Garza and Rouhana 1995).

Machine learning algorithms are the more advanced alternative to make the cost prediction. For instance, Hegazy and Ayed (1998) employed the neural network technique to develop a parametric cost estimating model for highway projects. Their study used 18 highway projects to develop neural network models through three optimization algorithms—back-propagation, simplex optimization, and generic—and compare their accuracies. The authors showed that the neural network model with simplex optimization provides higher accuracy in forecasting the construction costs than the other two algorithms. A study conducted by Al-Tabtabai et al. (1999) used a neural network to predict the preliminary cost of highway construction projects. The authors used location factors, project-participant factors, and project characteristics in developing a neural network model and showed the ability of the neural network technique to predict the cost of a highway project. Sodikov (2005) proposed an artificial neural network

(ANN) approach for developing cost estimates for highway projects during the conceptual phase. The author developed ANN models using project-related variables, such as project activities and project duration, and showed that the ANN model was superior in forecasting cost estimates for highway projects compared to the multiple regression model.

Chou (2009) examined the practicality of the case-based reasoning (CBR) technique for improving the accuracy of early cost prediction for pavement projects. The author concluded that the CBA model with experience-based weights of the attributes showed better predictability than the CBA model that contains attributes equally treated. Cheng et al. (2010) proposed an evolutionary fuzzy hybrid neural network, integrating neural networks and high order neural networks into a hybrid neural network, to estimate construction costs during the early stages of a project. The authors concluded that the proposed approach could yield better accuracy for construction conceptual cost estimates compared to the accuracy of the traditional neural network connections.

Petroutsatou et al. (2012) developed neural network models for predicting early cost estimates of road and tunnel construction projects. Their study showed the applicability of neural networks in developing forecasting models by capturing nonlinear data relationships. Gardner et al. (2016) developed cost estimating models using artificial neural networks and multiple regression techniques for highway projects. The study found that adding more variables does not necessarily increase the accuracy of cost estimates, and acceptable performance of the estimating models could be yielded using suitable input variables with low data collection and storage efforts. Adel et al. (2016) developed parametric models for conceptual cost estimates for highway projects using artificial neural networks. The authors used 75 highway projects and project-related factors, such as project duration, project region, and mainline length, in developing conceptual cost estimating models for highway projects. The authors concluded that the developed neural network models provide reliable accuracy for forecasting conceptual cost estimates for highway projects.

Cao et al. (2018) introduced the ensemble learning method in forecasting unit price bids for highway projects, which comprised four machine learning algorithms, including gradient boosting, extreme gradient boosting, random forest, and neural network. Their study showed the higher prediction accuracy of the proposed ensemble model compared to the accuracies of other methods, such as Monte Carlo simulation and multiple regression. In a follow-up work (Cao and Ashuri 2020), an advanced deep learning algorithm, long short-term memory (LSTM) algorithm, was used to predict highway construction costs. The authors showed that the proposed LSTM model outperformed the time series models in all three forecasting scenarios: short-term, medium-term, and long-term prediction. Moreover, Tijanić et al. (2019) used an artificial neural network to develop cost estimating models for road projects. The authors selected three neural networks—multilayer perceptron, generalized regression neural network, and radial basis function neural network—and compared the model accuracies. Their study showed that the neural network technique is a promising approach to estimate construction costs in the initial design phase of a project when there is a limited or incomplete set of data available. According to the literature, it can be concluded that machine learning algorithms have potential to improve data fitting to the forecasting models and provide higher accuracy in forecasting construction costs.

RESEARCH OBJECTIVE

Previous studies have shown that higher accuracy and better performance can be achieved by using machine learning algorithms to forecast construction costs. However, few studies have attempted to estimate the prices of the lump sum pay items for highway projects. One of the main challenges in developing accurate cost estimates for LS pay items (e.g., Traffic Control and Grading Complete) during PDP lies in the fact that LS pay items do not include individual items of work or quantities that constitute the LS measurement and payment, which can result in significant uncertainty in cost estimates of LS pay items. In addition, as project development advances, a vast amount of project information is generated and documented in several different databases, which makes it difficult for cost estimators to consider all the collected information items and quantify the information in developing cost estimates for LS pay items. The estimator must acquire better and more definite project information to develop an accurate cost estimate.

The overarching objective of this research is to develop forecasting models to estimate the prices of Traffic Control and Grading Complete LS pay items using advanced text mining and machine learning algorithms that detect key patterns of information generated during project development and provide higher accuracy of cost estimates.

CHAPTER 3. RESEARCH METHODOLOGY

INTRODUCTION TO DATA

This section describes data sources, including response variables and potential variables, for forecasting LS item prices. The prices of the Traffic Control and Grading Complete LS pay items used in highway construction projects in the state of Georgia were collected from the GDOT Cost Estimation System (CES). The research team collected prices of the Traffic Control LS pay items used in 304 highway projects and prices of the Grading Complete LS pay items used in 265 highway projects.

A Traffic Control LS item represents the work of managing mobility and safety impacts within a project work zone and addressing traffic safety and control through the work zone using items such as channelizing devices, temporary barriers, signage, traffic signals, warning devices, and pavement markings. A Grading Complete LS item represents the work of clearing, grubbing, and earthwork, including removals and excavating of all materials (e.g., ditches and undesirable materials), hauling, forming embankments, construction subgrades, etc.

The prices of the Traffic Control and Grading Complete LS pay items used in this research project were developed by the designer during project development at the milestone of the Final Field Plan Review (FFPR) cost estimate. Figure 1 shows the milestones of GDOT's project cost estimates during the plan development process.

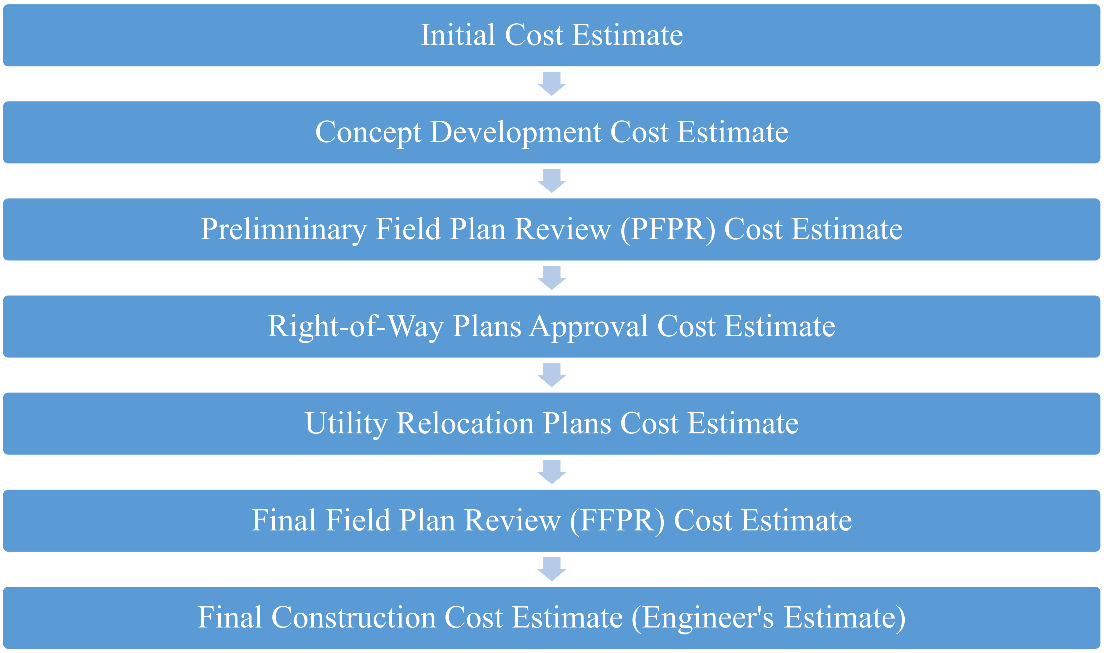

Initial Cost Estimate
Concept Development Cost Estimate
Prelimninary Field Plan Review (PFPR) Cost Estimate
Right-of-Way Plans Approval Cost Estimate
Utility Relocation Plans Cost Estimate
Final Field Plan Review (FFPR) Cost Estimate
Final Construction Cost Estimate (Engineer's Estimate)

Figure 1. Flow diagram. Milestones for project cost estimates during PDP.

In addition, the project development documents were collected from the GDOT databases to gather potential input attributes in developing a forecasting model. The project development documents include concept reports, final field plan review reports, and preconstruction status reports (PSRs). A concept report for a project is developed during the concept stage in coordination with subject matter experts (e.g., personnel in Right-of-Way, Utility, and Environmental Offices). Concept reports contain critical information related to a project, including project justification, project background, location of environmental resources, public involvement plan, access control, etc. During the final design stage, FFPR reports are developed to confirm that the design has efficiently and continuously satisfied the purpose and need of a project. FFPR reports summarize the review of plans and specifications, special provisions, permits, right-of-way agreements, and utility conflict resolutions for a project. Finally, during project development, the GDOT project manager uses preconstruction status reports to make critical decisions on a project. PSRs contain general project information, such as project location, project length, and work type, and track major tasks of project development and important notes regarding structure, right-of-way, utilities, etc. Table 1 shows the potential factors collected in this research effort that might impact the prices of the Traffic Control and Grading Complete LS pay items.

Table 1. Potential factors affecting Traffic Control and Grading Complete prices for highway projects.

Sources	Variables (66)	Descriptions	Units
Concept Reports	Construction Costs	Construction cost including construction, 5% Engineering and Inspection, Contingencies, and Liquid AC Cost Adjustment (not including ROW, Reimbursable Utility, PE, Environmental Mitigation Costs)	$
	Major Structure	Existence of major structures, such as bridges and retaining walls	Boolean Indicator
	Major Interchange	Existence of major interchanges in the project location	Boolean Indicator
	Major Intersection	Existence of major intersections in the project location	Boolean Indicator
	Construction Issues Potentially Affecting Constructability/ Construction Schedule	Existence of potential issues that affect constructability or construction schedules	Boolean Indicator
Final Field Plan Review Reports	Current Traffic Average Daily Traffic (ADT)	The total volume of vehicle traffic	Number
	Number of Parcels for Right of Way	Number of parcels for the right-of-way required in a project	Number
	Estimated Contract Time	Estimated contract duration for a project	Month
	Types of Traffic Control Plans	Seven types of traffic control plans, including: • Detours • Lane Closures • Lane Closures and Detours • Lane Closures, Detour, and Flagging Operations • Lane Closures and Flagging Operations • Traffic Restrictions • No Traffic Restrictions	Boolean Indicator
	Summary of Quantities (06-XXX)	Presence of words related to traffic control (3): • Temporary barrier • Attenuator • Striping	Boolean Indicator
	Comments on Mainline Roadway Plan Sheet (13-XXX)	Presence of comments for Mainline Roadway Plan Sheet (13-XXX) in the Field Plan Review Reports	Boolean Indicator

	Comments on Construction Staging & Cross-Section Plan Sheet (19-XXX)	Presence of comments for Construction Staging & Staging Cross-section Plan Sheet (19-XXX) in the Field Plan Review Reports	Boolean Indicator
		Presence and occurrence of words related to traffic control (4): • Stage • Shift • Closure • Detour	Number
	Comments on Drainage Profiles (22-XXX)	Presence of comments for Drainage Profiles (22-XXX) in the Field Plan Review Reports	Boolean Indicator
	Comments on Retaining Walls Envelopes (31-XXX)	Presence of comments for Retaining Walls Envelopes (31-XXX) in the Field Plan Review Reports	Boolean Indicator
	Comments on Retaining Walls Plans (32-XXX)	Presence of comments for Retaining Walls plans (32-XXX) in the Field Plan Review Reports	Boolean Indicator
	Comments on Bridge Plans (35-XXX)	Presence of comments for Bridge Plans (35-XXX) in the Field Plan Review Reports	Boolean Indicator
	Project Type	Highway project types (8): • Widening & Passing Lanes • New Location Roadways • Interchange Reconstructions • Grade Separations • Bridge Program (e.g., replacement of an existing bridge or construction of a bridge where there is no existing bridge) • Intersection Improvements (e.g., roundabouts, signals, other intersection control type changes) • Other Operational Improvements (e.g., pedestrian upgrade, lighting, advanced traffic management (ATM), information technology services (ITS), and other operational improvements • Systematic Improvements: Improvements of guardrail, cable barrier, drainage, and noise wall	Boolean Indicator
	Functional Classification	Functional classification of the project (24): • Pedestrian Facility • Rural Freeway • Rural Major Arterial • Rural Major Collector • Rural Major Principal Arterial • Rural Minor Arterial • Rural Minor Collector • Rural Minor Interstate Principal Arterial • Rural Minor Local Road • Rural Minor Principal Arterial • Urban Freeway and Expressway • Urban Major Arterial • Urban Major Collector • Urban Major Interstate • Urban Major Interstate Principal Arterial • Urban Major Principal Arterial • Urban Minor Arterial • Urban Minor Collector • Urban Minor Interstate • Urban Minor Interstate Principal Arterial • Urban Minor Local Road • Urban Minor Principal Arterial • Varies • Not Provided	Boolean Indicator
	Major Project	A project that has significant amounts of right-of-way acquisition; a significant change in travel patterns; or significant social, economic, or environmental effects	Boolean Indicator

Preconstruction Status Reports	Road Type	The types of roadways based on the project description in the PSR (3): - County Road - State Route - US Route Interstate	Boolean Indicator
	Project Length	Length of the project	Miles
	Metropolitan Planning Organizations (MPOs)	Areas with a population greater than 50,000, defined by the U.S. Census	Boolean Indicator

To develop the forecasting models for the prices of the Traffic Control and Grading Complete LS pay items, 66 initial attributes were collected from the GDOT project development documents, including concept reports, FFPR reports, and PSRs.

OVERVIEW OF THE RESEARCH METHODS

Figure 2 illustrates an overview of the proposed methodology employed in this paper. The research methodology includes four stages: data processing, feature selection, model selection, and final model development. Multiple steps of machine learning algorithms are introduced to achieve the objective of this research.

First, the researchers used a text-mining algorithm to convert text information into quantifiable data and a dimension reduction algorithm to reduce the computational complexity of the model (Ramsingh and Bhuvaneswari 2021). Since project scope and design documents are stored with unstructured text files, a text-mining algorithm, term frequency–inverse document frequency (TF–IDF), enabled categorization of the text data and index words per their importance in the document by identifying the frequency of words and checking the presence of the words related to the Traffic Control and Grading

Complete LS pay items. TF–IDF has been used widely for analyzing construction (ul Hassan et al. 2020, Moon et al. 2021, Jafari et al. 2021). In addition, a dimension reduction algorithm, principal component analysis (PCA), was utilized to determine the optimal reduced feature set, which can yield the optimal accuracy and learning times. As the information in the project documents is vast, the dimension reduction algorithm reduced the dimensionality of words identified from the text mining algorithm and transformed them into a new set of features (Lin et al. 2015).

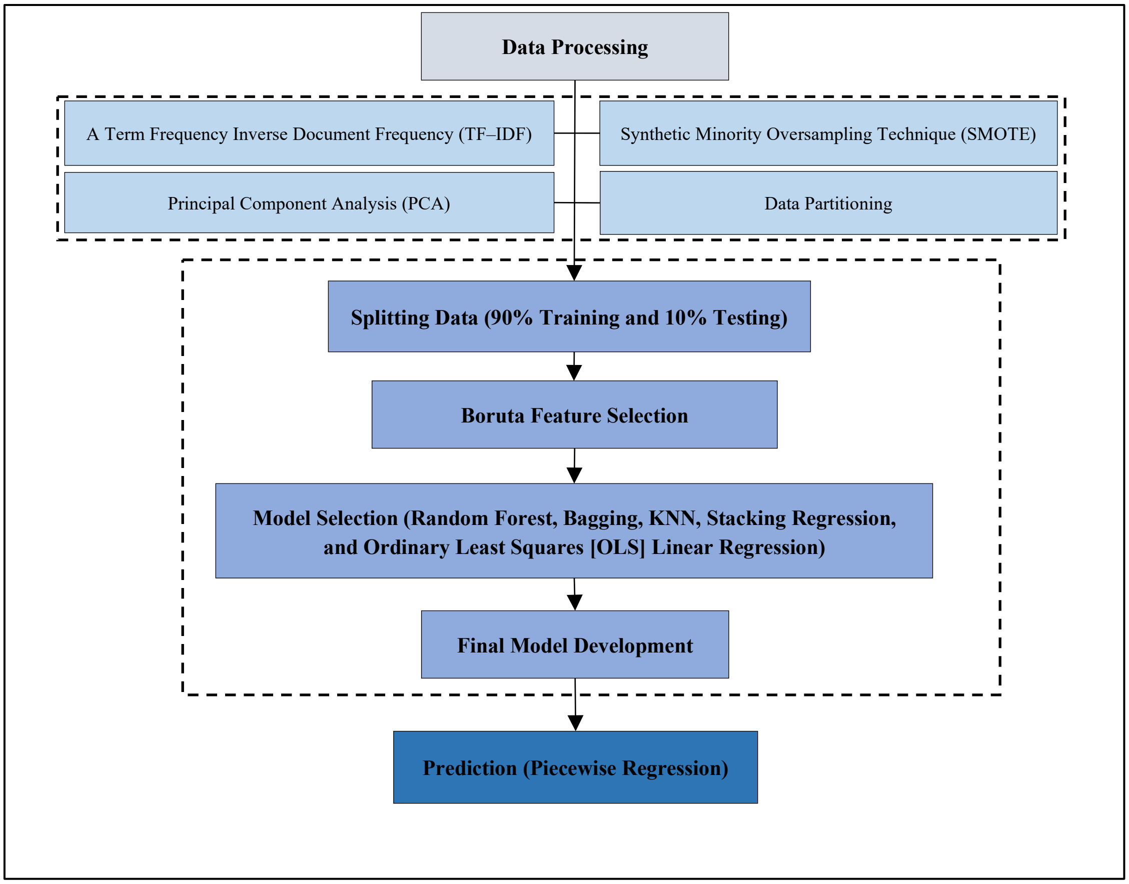

Data Processing
A Term Frequency Inverse Document Frequency (TF–IDF)
Synthetic Minority Oversampling Technique (SMOTE)
Principal Component Analysis (PCA)
Data Partitioning
Splitting Data (90% Training and 10% Testing)
Boruta Feature Selection
Model Selection (Random Forest, Bagging, KNN, Stacking Regression, and Ordinary Least Squares [OLS] Linear Regression)
Final Model Development
Prediction (Piecewise Regression)

Figure 2. Flow diagram. Overview of research methodology.

Second, this research used an oversampling algorithm—synthetic minority oversampling technique (SMOTE)—to overcome the issue of imbalanced data. Imbalanced data or skewed data sets are challenging because they will cause bias for developing a forecasting model. For instance, there will be high predictive accuracy for the majority data while the minority data will have poor predictivity because the minority class is treated as noise and ignored completely by the classifier (Su and Hsiao 2007). Thus, SMOTE will improve classifier performance for the minority class. Next, data partitioning was conducted to minimize the errors of forecasting models by making the predicted values of the forecasting models close to the observed data points. This research identified data segments by examining a quantile plot of construction costs. And then, the data were divided into training and testing datasets, 90 and 10 percent, respectively.

Next, the BORUTA feature selection algorithm was used to select important variables to predict the prices of the Traffic Control and Grading Complete LS pay items. Since including a higher number of variables for developing a forecasting model can decrease the efficiency of the algorithm and the accuracy of a forecasting model (Kohavi and John 1997), identifying a set of important variables was crucial. This research used the Boruta feature selection algorithm because it provides an unbiased and stable selection of important and non-important variables (Kursa and Rudnicki 2010). With identified variables for data segments from the Boruta algorithms, multiple machine learning algorithms, including random forest, bagging, KNN, stacking regressor, and ordinary least squares (OLS) linear regression, were used to develop forecasting models for data segments. For the data segment, this research selected the best algorithms that provide the least absolute error. Finally, using the best algorithm for each data segment, piecewise regression was employed to predict the prices of the Traffic Control and Grading Complete LS items.

DATA PROCESSING

Text Mining with Term Frequency–Inverse Document Frequency and Principal Component Analysis

The forecasting accuracy heavily depends on the data quality used for developing forecasting models. Data processing was carried out in the following steps. First, useful information hidden in collected text documents (i.e., concept reports and FFPR reports) was extracted using the term frequency–inverse document frequency technique, which allows for ranking of important words within documents and creates a document term matrix based on the importance of the identified words. Raw counts of specific words were also computed from the text information. To extract important words related to the Traffic Control and Grading Complete LS pay items, this research applied the TF–IDF technique to several text components in concept reports and FFPR reports, as presented in Table 2 and Table 3.

Table 2. Text components for Traffic Control LS pay item.

Data Source	Text Components
Concept Reports	Need and purpose project justification
FFPR Reports	Project Description
	Special Provisions General Special Provision
	Typical sections
	Summary of Quantities
	Construction Staging & Cross-Section Plan Sheet (19-XXX)
	Construction Issues potentially affecting constructability/construction schedule

Table 3. Text components for Grading Complete LS pay item.

Data Source	Text Components
Concept Reports	Need and purpose project justification
FFPR Reports	Project Description
	Cover Sheet (01-XXX)
	Mainline Roadway Plan Sheet (13-XXX)
	Cross Sections (23-XXX)
	Summary of Quantities
	Construction Staging & Cross-Section Plan Sheet (19-XXX)

As the developed document term matrix from the TF–IDF technique is large and sparse, principal component analysis was applied to reduce feature dimensions (i.e., document term matrix). PCA uses a linear combination of initial text variables identified from the TF–IDF (Kotekar and Kamath 2018) to reduce dimensionality through the orthogonal projection of the original data onto a lower-dimensional linear space. The identified principal components transform the textual variables describing the project into an optimal subspace serve. The calculated principal components are valuable as intermediate links for the subsequent forecasting analysis.

Synthetic Minority Oversampling Technique and Data Partitioning

A synthetic minority oversampling technique was introduced to address highly right-skewed data. Figure 3 and figure 4 depict the distribution of prices of the Traffic Control and Grading Complete LS pay items, which are highly right-skewed. SMOTE combines undersampling of the frequent data points with oversampling of the minority data points, which reduced skewness for the collected data of the Traffic Control and Grading Complete prices. SMOTE has the potential to improve the accuracy of model performance by handling the skewed class distribution. To oversample the minority class, synthetic samples were generated by interpolating between adjacent minorities. The sampling was performed in the same feature space of original minority classes, which allows for more consistent data distribution (Rong et al. 2014). The majority class was undersampled by randomly removing samples from the majority class population until the minority class became some specified percentage of the majority class. Thus, SMOTE used the combination of undersampling and oversampling to address data skewness (Chawla et al. 2002).

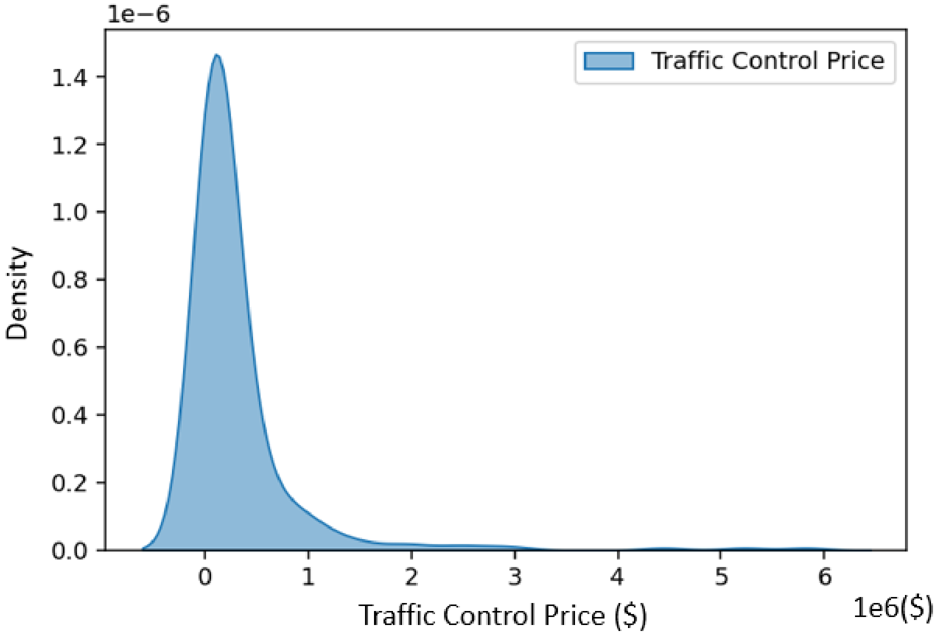


Figure 3. Graph. Distribution of Traffic Control price.

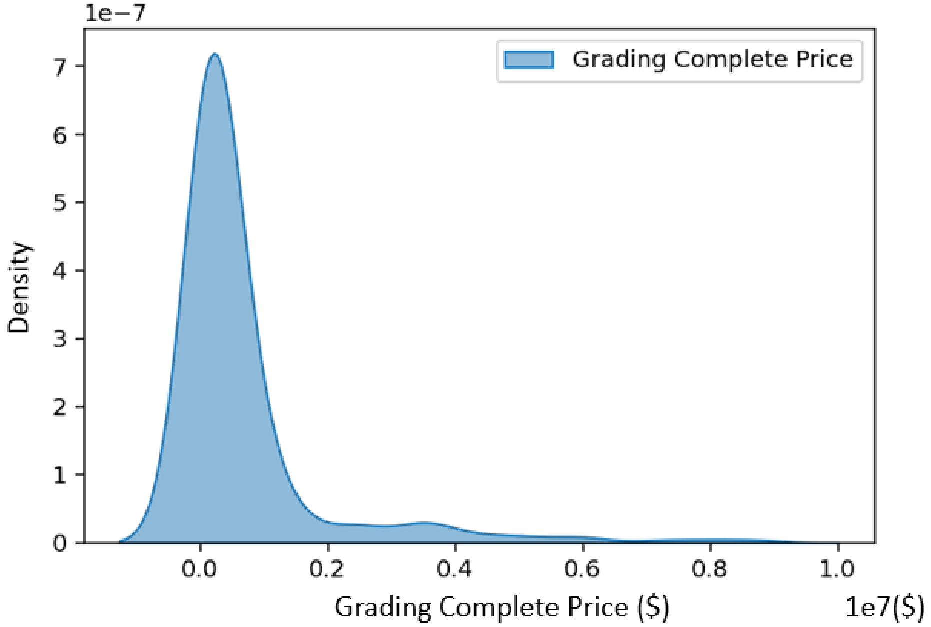


Figure 4. Graph. Distribution of Grading Complete price.

Feature Selection with Boruta Feature Selection Method

The primary purpose of feature selection is to select a subset of variables from the potential variables that are useful to explain the responsible variable while excluding effects from irrelevant and redundant variables (Guyon and Elisseeff 2003).

This research used a Boruta feature selection technique to select the most relevant variables with the greatest potential for enhancing price prediction of Traffic Control and Grading Complete LS items. Following Kursa and Rudnicki's (2010) recommendations, the following steps were implemented to find relevant variables for developing forecasting models. First, the Boruta algorithm extended the information system by adding copies of all variables (e.g., adding at least five shadow attributes). The added attributes were randomly shuffled to remove their correlations with the response variable. Next, the algorithm ran a random forest algorithm on the extended information system and gathered the Z scores computed. The maximum Z score among shadow attributes (MZSA) was determined as the threshold. A two-sided test of equality with the MZSA was performed for each attribute with undetermined importance. The algorithm deemed the attributes with importance significantly higher than MZSA as important variables and those with importance significantly lower than MZSA as unimportant variables. Lastly, unimportant attributes were permanently removed from the information system. The Boruta feature selection technique has proven successful in prior studies to identify the sets of variables for developing a forecasting model (Cao et al. 2018, Assaad and El-adaway 2020).

MACHINE LEARNING ALGORITHMS

Multiple machine learning algorithms, including random forest, bagging, KNN, stacking regressor, and OLS linear regression, were used in this research to develop forecasting models for prices of Traffic Control and Grading Complete LS pay items and to identify the best performance models for each data segment. Using the machine learning algorithms, forecasting models were developed for each defined segment, and their performances were compared through the mean absolute percentage error.

Random Forest

The random forest is one of the most accurate machine learning algorithms in terms of reducing bias and overfitting. The forest makes trees by randomly taking observations and input variables to decorrelate the base learners (Nasiriany et al. 2019, Murphy 2012). It aggregates a bundle of decision trees and takes the total average for prediction. The trees create branches by learning the decision rules inferred from the data. Each branch contains a set of rules and selected features that correspond to an output at the end of the branch. The primary advantage of this algorithm is the capability to consider both categorical and continuous variables in developing a forecasting model.

Bagging

The “bagging” algorithm is an ensemble estimator composed of bootstrap and aggregation (Elmousalami 2020). The bootstrap step randomly draws replicas from the training dataset with replacement. Base regressors are fit on each random subset. The aggregation step forms the final prediction by averaging outcomes from base regressors (Breiman 1996). The algorithm leverages predictions from weak models that specialize in a different dimension of feature space to improve accuracy, reduce variance, and eliminate the chance of overfitting.

K-nearest Neighbors

The k-nearest neighbors regression is a nonparametric method that uses feature similarity to produce an estimation for new data entries. It is reasonable to conjecture that those similar observations should be nearby (Song et al. 2017). The algorithm approximates the relationship between the input features and the outcome by local interpolation of observations from the same neighborhood. The inverse distance is used to assign weight to different training samples.

Stacking Regressor

An ensemble learning technique is used to stack the outputs from individual regressors through ridge regression. The outputs from the random forest, bagging, and KNN are taken as inputs of the stacking regressor, allowing for higher generalizability. Ridge regression is a natural fit as it solves the problem of overfitting, especially when the data suffer from multicollinearity (Mohammadi 2020). The algorithm formulation shrinks the parameters by pushing the coefficients toward zero. Such regularization reduces model complexity and the risk of multicollinearity arising in the unseen data.

Model Evaluation

The measure of mean absolute percentage error was used in this research to investigate the model accuracy on testing data. The MAPE evaluates the relative error of the forecasting models, as well as combines approaches using a stacking regressor. Because the test set is not used for model fitting, it reflects the model predictability when handling unknown future events. The lower the MAPE value, the higher the accuracy of forecasting models; the higher the MAPE value, the lower accuracy of forecasting models. The metric is defined in equation 1 below.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|e_i|}{y_i} \times 100$$

(**Error! Bookmark not defined.**1)

where $\boldsymbol{y_i}$ is the actual Traffic Control or Grading Complete prices of project *i*, $\boldsymbol{e_i}$ is the residual of project *i*, and *n* is the number of data points that are predicted.

CHAPTER 4. DATA COLLECTION AND PROCESSING

RESULTS OF DATA PROCESSING

Critical terms were identified, and the numerical representations of text data were developed through a term frequency–inverse document frequency technique. The PCA technique was performed on the developed sparse document term matrix to reduce dimensionality through the orthogonal projection of the original data onto a lower-dimensional linear space. The first 21 principal components were kept after PCA. Finally, multicollinearity among input variables was detected using the variance inflation factor (VIF). If the value of the VIF is greater than 10 for a variable, severe multicollinearity exists in the model. The following four variables were removed from the subsequent modeling process:

- Estimate Contract Time.
- Presence of Word "Stage" in Construction Staging & Cross-Section Plan Sheet (19-XXX).
- Presence of Comments on Mainline Roadway Plan Sheet (13-XXX).
- Project Type Bridge Program.

Therefore, this research identified 81 potential variables in developing forecasting models for the prices of the Traffic Control and Grading Complete LS pay items.

Next, SMOTE was applied to the collected prices of the Traffic Control and Grading Complete LS pay items to address highly right-skewed data. Data partitioning was implemented to minimize the errors, which allows for making the predicted values close to the observed data points. Furthermore, data partitioning is helpful to capture the different relationships between the response variable and independent variables in different partitions, especially when the response variable spans an extensive range. A quantile plot of construction costs for highway projects is used to determine breakpoints of data segments. Figure 5 presents a quantile plot of construction costs for highway projects, which were collected for the Traffic Control LS pay item. Figure 6 depicts a quantile plot of construction costs for highway projects, collected for the Grading Complete LS pay item.

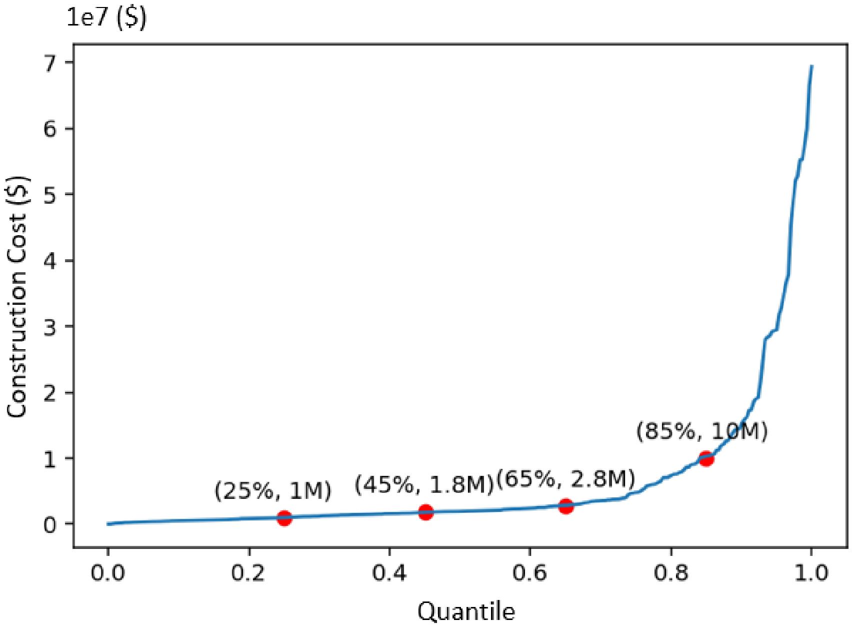


Figure 5. Graph. Quantile plot of construction cost for Traffic Control LS pay item.

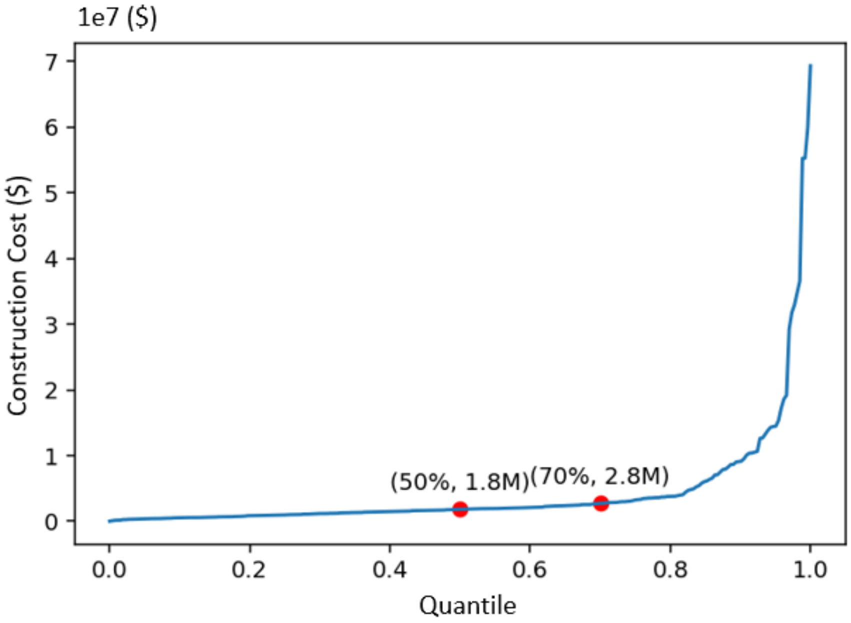


Figure 6. Graph. Quantile plot of construction cost for Grading Complete LS pay item.

Table 4 provides breakpoints for data segments of the Traffic Control and Grading Complete LS pay items. The results in table 4 indicate that five segments were determined for the Traffic Control LS pay item and three segments were determined for the Grading Complete LS pay item.

Table 4. Results of data partitioning.

Segments	Traffic Control	Grading Complete
	Interval (Construction Costs $)	
Segment 1	<$1Million	<$1.8Million
Segment 2	[$1 Million, $1.8 Million)	[$1.8 Million, $2.8 Million)
Segment 3	[$1.8 Million, $2.8 Million)	>=$2.8 Million
Segment 4	[$2.8 Million, $10 Million)	–
Segment 5	>=$10 Million	–

Finally, the data were divided into training and testing data sets for model development and validation. This research uses 90 percent of the randomly selected data for training the model and the remaining 10 percent for testing the model's predictability.

RESULTS OF FEATURE SELECTION

For each identified segment, the Boruta algorithm was applied to select the most important variables and remove unimportant variables to improve the model parsimony. The results of a Boruta feature selection for the Traffic Control and Grading Complete LS pay items are presented in Table 5 and Table 6, respectively.

Table 5. Results of Boruta feature selection for Traffic Control LS pay item.

Ranks	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5
	<$1M	[$1M, $1.8M)	[$1.8M, $2.8M)	[$2.8M, $10M)	>=10M
	Features				
1	Construction Cost	Traffic ADT	Construction Cost	Project Length	Project Length
2	State Route	Construction Cost	Principal Component 8	Number of Parcels for ROW	Construction Cost
3	Not Provided (Functional Classification)	Number of Parcels for ROW	Project Length	Construction Cost	Traffic ADT
4	Stage#	Project Length	State Route	Traffic ADT	Number of Parcels for ROW
5	Systematic Improvements	Rural Minor Local Road (Functional Classification)	Number of Parcels for ROW	State Route	No Traffic Restrictions & Control Plans
6	Project Length	Striping	Traffic ADT	Rural Minor Interstate Principal Arterial (Functional Classification)	Other Operational Improvements
7	Comments on Drainage Profiles	Rural Major Collector (Functional Classification)	Striping	Stage#	Stage#
8	Traffic ADT	Stage#	Lane Closures	Comments on Retaining Walls Envelopes	Intersection Improvements
9	Lane Closures	Principal Component 4	Other Operational Improvements	Closure#	Detour#
10	Number of Parcels for ROW	Intersection Improvements	Rural Major Collector (Functional Classification)	Comments on Bridge Plans	Urban Minor Principal Arterial (Functional Classification)
11	MPOs	New Location Roadways		Striping	Shift#
12	County Road				Urban Major Principal Arterial (Functional Classification)
13	Comments on Bridge Plans				Comments on Major Structure
14					Urban Major Interstate (Functional Classification)

According to the results for a Traffic Control LS pay item in Table 5, 13 variables were selected in Segment 1. Construction cost, state route, no functional classification, and word frequencies of stages in the field plan review reports were the most critical factors in Segment 1, followed by systematic improvements, project length, and binary variable of comments on drainage profiles. In Segment 2, the results indicated that 11 variables were selected. Traffic volume, construction cost, number of parcels for ROW, and project length were the most significant variables, followed by rural minor local road (functional classification), striping (summary of quantity), and rural major collector (functional classification). In Segment 3, 10 variables were selected. The results showed that construction cost, principal component 8, project length, and state route were the most important variables, followed by the number of parcels for ROW, traffic volume, and striping. The identified principal component points to one of the principal directions along which the text data show the largest variation. In Segment 4, 11 variables were selected. The results indicated that project length, number of parcels, construction cost, and traffic volume were the most important variables, followed by state route, rural minor interstate principal arterial, word frequencies of the stage, the binary variable of comments on retaining walls envelopes, and word frequencies of closure. Lastly, Segment 5 had 14 variables identified. The most important variables in Segment 5 were project length, construction cost, traffic volume, and the number of parcels for ROW, followed by no traffic restrictions & control plans (type of traffic control plan), other operational improvements, word frequencies of stages, and intersection improvements.

Furthermore, the results showed that four variables—construction cost, project length, traffic volume, and the number of parcels—were influential in all segments.

Table 6. Results of Boruta feature selection for Grading Complete LS pay item.

Ranks	Segment 1 (<$1.8M)	Segment 2 [$1.8M,$2.8M)	Segment 3 (>=$2.8M)
	Features		
1	Construction Cost	Traffic ADT	LENGTH_MI
2	Principal Component 4	LENGTH_MI	Stage#
3	LENGTH_MI	Construction Cost	Construction Cost
4	Principal Component 17	Principal Component 9	Traffic ADT
5	Traffic ADT	ROW Number of Parcels	ROW Number of Parcels
6	Stage#	Principal Component 18	FC Urban Major Principal Arterial
7	Principal Component 19	Principal Component 5	Principal Component 13
8	Comments on Major Interchanges	Principal Component 12	Principal Component 7
9	Principal Component 18	Principal Component 8	Principal Component 3
10	Principal Component 6	Principal Component 14	Principal Component 8
11	Principal Component 11	Principal Component 19	Principal Component 1
12	Principal Component 1	Principal Component 0	Principal Component 19
13	Principal Component 12	Principal Component 16	Principal Component 16
14	FC Not Provided	Principal Component 20	Principal Component 15

Table 6 provides the results of feature selection for the Grading Complete LS pay item; 14 variables were selected in Segments 1, 2, and 3, respectively. In Segment 1, construction cost, principal component 4, project length, principal component 17, and traffic volume are the most critical factors, followed by the number of occurrences of word "stage," principal component 19, comments on major interchanges, etc. In Segment

2, the results indicated that traffic volume, project length, construction cost, principal component 9, and the number of parcels for ROW are the most significant variables, followed by other principal components. The selected principal components represent the directions that explain the largest amount of variation in the text documents. For Segment 3, the results showed that project length, the number of occurrences of the word "stage," construction cost, traffic volume, and the number of parcels for ROW are the most important variables, followed by the functional classification of urban major principal arterial and other principal components. Overall, the results showed that three variables—construction cost, project length, and traffic volume—were significantly influential in all segments for predicting prices of a Grading Complete LS pay item. Thus, the identified variables for each segment for the Traffic Control and Grading Complete LS pay items were used for developing forecasting models.

CHAPTER 5. DEVELOPMENT OF FORECASTING MODELS FOR LUMP SUM PAY ITEMS

OVERVIEW

The prices of major lump sum pay items, Traffic Control 150-1000 and Grading Complete 210-0100, were forecasted using key project attributes identified through data processing and feature selection. Five machine learning algorithms were employed: KNN, random forest, bagging, stacking regressor, and OLS linear regression to train models for each data segment using 90 percent of the dataset. The out-of-sample performance of models was evaluated using MAPE on the remaining testing data. The following sections present evaluations of the forecasting results in both tabular and graphic formats.

TRAFFIC CONTROL LUMP SUM ITEM (150-1000)

The following subsections examine the model performance from three aspects: model comparison, overall accuracy, and residual examination.

Model Comparison

Table 7 presents the out-of-sample MAPE values for predicting the Traffic Control LS item price using different algorithms for each segment. In Segment 1, KNN gave the most accurate prediction, compared to the other algorithms. The MAPE value for KNN in Segment 1 is 6.25 percent. In Segment 2, random forest was the best-performing algorithm in terms of MAPE. In Segment 3, KNN showed the best performance, compared to the other algorithms; the MAPE value for KNN in Segment 3 is 8.04 percent. The random forest algorithm demonstrated the highest accuracy (i.e., MAPE) in forecasting the Traffic Control LS pay item prices in Segment 4. For Segment 5, the stacking regressor algorithm showed the best accuracy in forecasting prices of the Traffic Control LS pay items. The MAPE value for stacking regressor in Segment 5 is 7.16 percent. Therefore, this study selected KNN, random forest, KNN, random forest, and stacking regressor for Segments 1, 2, 3, 4, and 5, respectively. It can also be noted that the selected machine learning algorithms for the five segments outperformed the ordinary least squares linear regression based on the MAPE values. Consequently, the results showed that the estimated accuracy of the developed machine learning models for the segments ranged from 6.25 to 14.05 percent.

Table 7. Test results of model selection for segments (Traffic Control).

Segments (Construction Costs $)	Models	MAPE (%)
Segment 1 **<$1M**	**K-Nearest Neighbors (KNN)**	**6.25**
	Random Forest	11.84
	Bagging	12.66
	Stacking Regressor (Ridge)	16.73
	OLS Linear Regression	18.89
Segment 2 **[$1M, $1.8M)**	**Random Forest**	**14.05**
	Bagging	21.79
	Stacking Regressor (Ridge)	25.72
	K-Nearest Neighbors (KNN)	39.82
	OLS Linear Regression	50.86
Segment 3 **[$1.8M, $2.8M)**	**K-Nearest Neighbors (KNN)**	**8.04**
	Random Forest	15.73
	Bagging	17.07
	Stacking Regressor (Ridge)	30.42
	OLS Linear Regression	36.83
Segment 4 **[$2.8M, $10M)**	**Random Forest**	**12.52**
	Stacking Regressor (Ridge)	12.74
	K-Nearest Neighbors (KNN)	13.44
	Bagging	16.07
	OLS Linear Regression	43.00
Segment 5 **>=$10M**	**Stacking Regressor (Ridge)**	**7.16**
	K-Nearest Neighbors (KNN)	12.87
	Random Forest	14.33
	Bagging	27.62
	OLS Linear Regression	27.90

Overall Accuracy

Next, the accuracy of the forecasting models was compared between partitioned data and data without partitioning (table 8). The overall accuracy of the proposed model was 90.21 percent (MAPE = 9.79 percent). It was concluded that the developed machine learning models achieve a high level of forecast accuracy. In contrast, the machine learning models achieve much lower accuracy when applied to data without partitioning.

The error metric of OLS reaches more than 100 percent. Therefore, the piecewise regression, which partitions data into different intervals and fits a regression function to each one, significantly improves the model predictability.

Table 8. Test results comparison between partitioned data and data without partitioning (Traffic Control).

	Partitioned Data	**Data without Partition**				
Algorithms	*Proposed Model*	*KNN*	*Stacking Regressor*	*Random Forest*	*Bagging*	*OLS*
MAPE (%)	9.79	28.53	31.68	36.88	45.69	109.37

Figure 7 shows the scatter plot of actual and predicted traffic control prices. Since the scatter plots are close to the red dashed line, the developed forecasting models show the strong capability to predict the prices of the Traffic Control LS items.

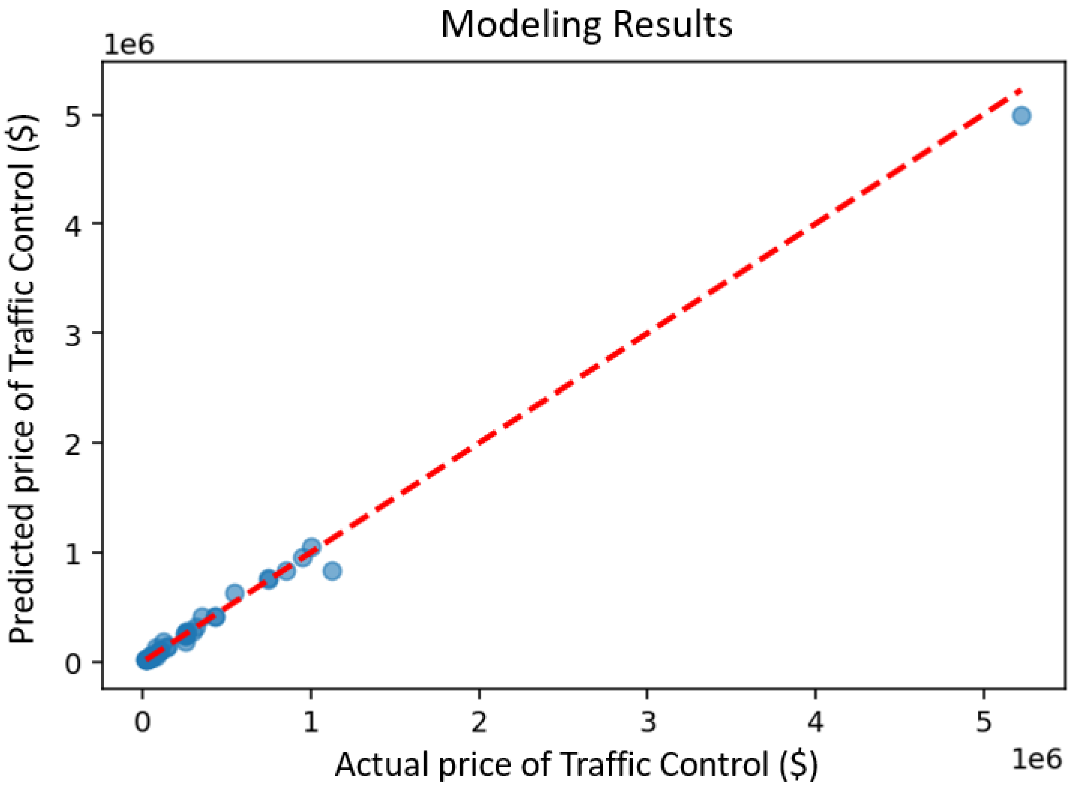


Figure 7. Graph. Model results of piecewise regression (Traffic Control).

Residual Examination

Finally, the residuals obtained from the forecasting models were examined using the scatter plot and D'Agostino's K-squared test. Figure 8 provides the scatter plot of residuals against fitted points for the Traffic Control LS pay item. Nearly half the data points are below the zero-line and half above the line. The plot also suggests no trend between residuals and fitted points.

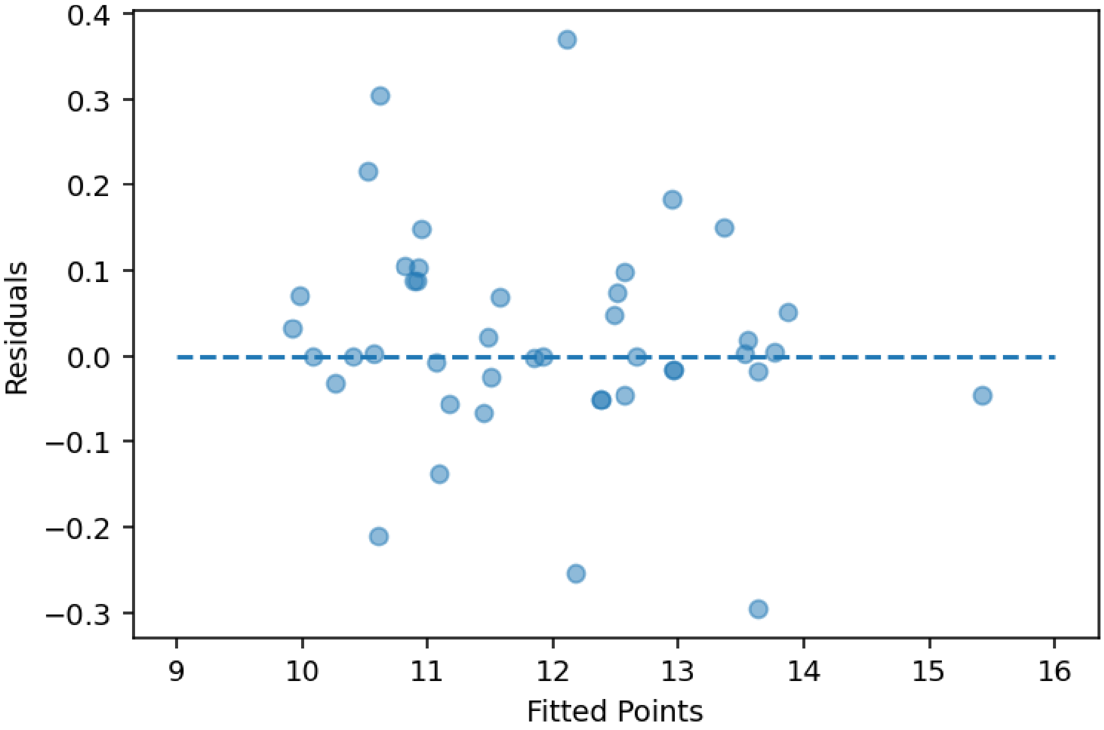


Figure 8. Graph. Scatter plot of residuals (Traffic Control).

The result of the D'Agostino's K-squared test, provided in table 9, indicates that the null hypothesis that the residuals follow the normal distribution is accepted at a significance level of 0.05. Therefore, it is clear that the forecasting models adequately capture the information in the data.

Table 9. Normality test of residuals (Traffic Control).

Test	Statistics	p-value
D'Agostino's K-squared test	5.049	0.080

GRADING COMPLETE LUMP SUM ITEM (210-0100)

The following subsections examine the model performance from three aspects: model comparison, overall accuracy, and residual examination.

Model Comparison

The out-of-sample MAPE values for forecasting the Grading Complete LS item price using different machine learning algorithms for each segment are summarized in table 10. For Segment 1 and Segment 2, KNN achieved the highest accuracy compared to the other machine learning algorithms, with a MAPE of 7.34 and 3.56 percent, respectively. For Segment 3, random forest stood out as the most accurate approach for forecasting the Grading Complete LS item price. Therefore, this study selected KNN, KNN, and random forest for Segments 1, 2, and 3, respectively. The out-of-sample MAPE values ranged from 3.56 to 14.31 percent. All the machine learning algorithms outperformed the ordinary least squares linear regression in terms of MAPE values.

Table 10. Test results of model selection for segments (Grading Complete).

Segments (Construction Costs $)	Models	MAPE (%)
Segment 1 <$1.8M	**K-Nearest Neighbors (KNN)**	**7.34**
	Random Forest	16.64
	Stacking Regressor (Ridge)	19.41
	Bagging	20.59
	OLS Linear Regression	57.20
Segment 2 [$1.8M, $2.8M)	**K-Nearest Neighbors (KNN)**	**3.56**
	Random Forest	13.03
	Stacking Regressor (Ridge)	13.70
	Bagging	22.51
	OLS Linear Regression	64.80
Segment 3 >=$2.8M	**Random Forest**	**14.31**
	Bagging	18.01
	Stacking Regressor (Ridge)	23.18
	K-Nearest Neighbors (KNN)	24.63
	OLS Linear Regression	99.08

Overall Accuracy

The overall accuracy was compared between the developed piecewise regression using partitioned data and machine learning algorithms using data without partitioning (table 11). The overall accuracy of the proposed model was 91.40 percent (MAPE = 8.60 percent), which was much higher than the other approaches that used data without partitioning. The MAPE values ranged from 37.05 to 147.51 percent when forecasting algorithms were directly applied to data without partitioning. The high accuracy of the proposed model validated the effectiveness of piecewise regression, which could capture the complex relationship between an LS item price and project features in different intervals.

Table 11. Test results comparison between partitioned data and data without partitioning (Grading Complete).

	Partitioned Data	Data without Partition				
Algorithms	*Proposed Model*	*Random Forest*	*Stacking Regressor*	*Bagging*	*KNN*	*OLS*
MAPE (%)	8.60	37.05	38.99	39.39	44.86	147.51

Figure 9 provides the graphic evaluation of the model performance. The points lie close to the 45-degree line, confirming the model predictability.

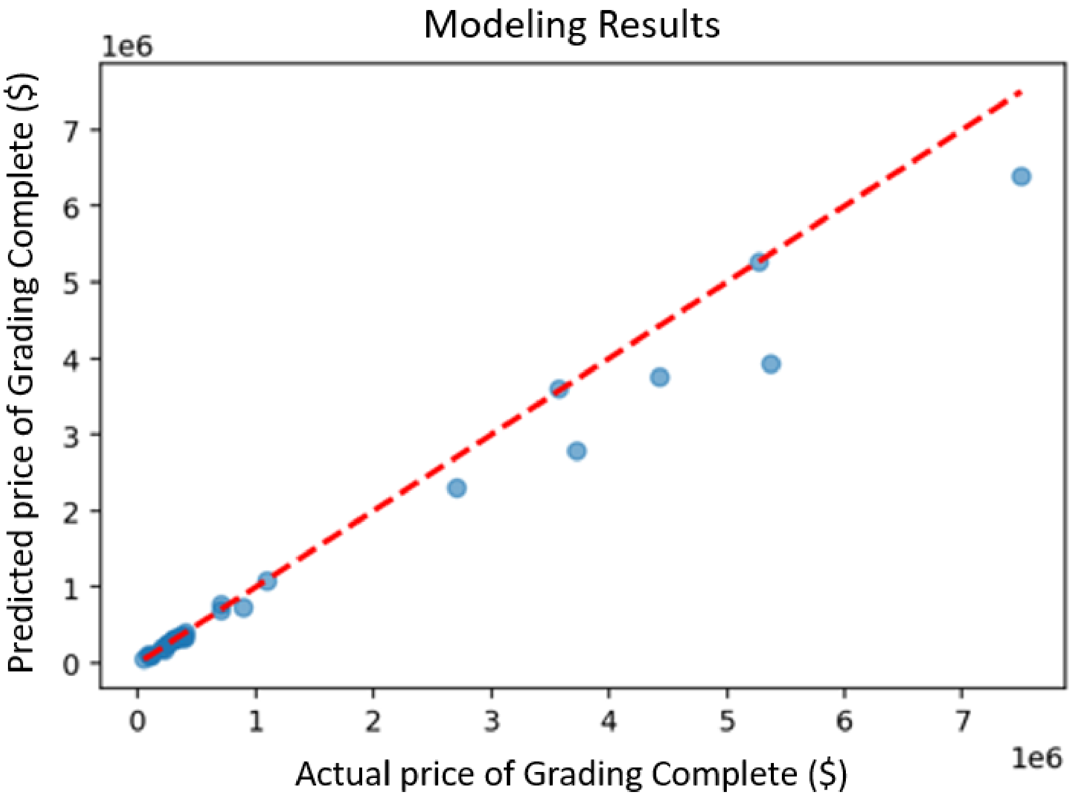


Figure 9. Graph. Model results of piecewise regression (Grading Complete).

Residual Examination

The residual examination was performed using a scatter plot and normality test. In figure 10, the residuals are randomly scattered around zero. The degree of scattering is constant for all fitted values. The residuals show no pattern against the fitted points, which suggests that the proposed model appropriately captured all the trends in the data.

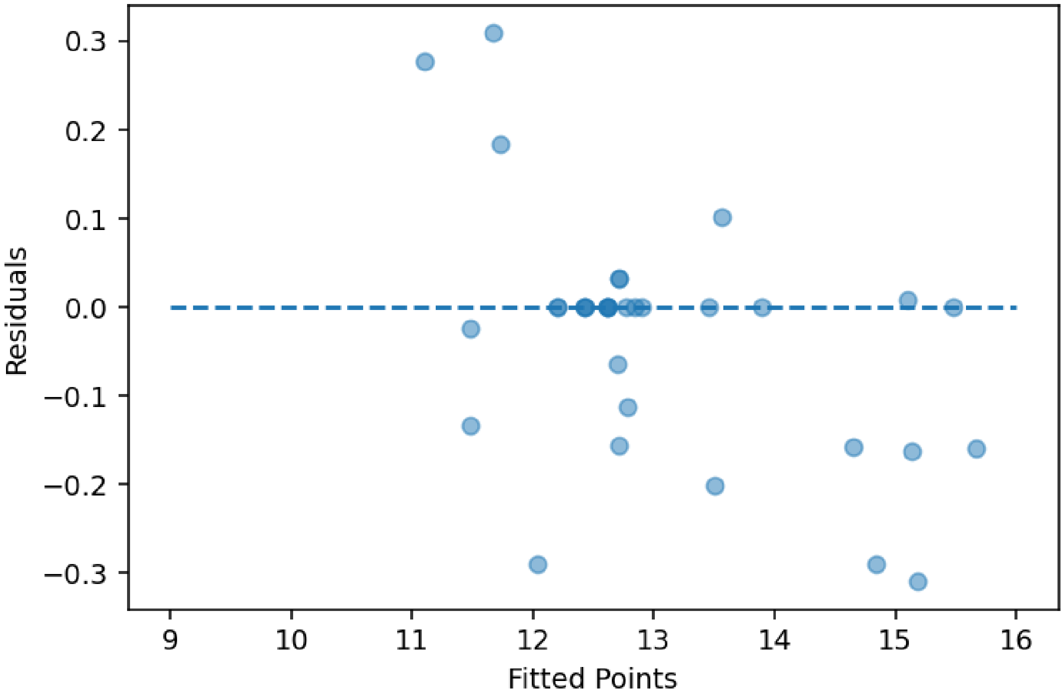


Figure 10. Graph. Scatter plot of residuals (Grading Complete).

The D'Agostino's K-squared test statistics also confirmed the residuals' normal distribution at a significance level of 0.05 (table 12). Therefore, it was concluded that the proposed model thoroughly described the relationship between the key project features and LS item price, and only random error remained.

Table 12. Normality test of residuals (Grading Complete).

Test	Statistics	p-value
D’Agostino’s K-squared test	1.884	0.390

CHAPTER 6. WEB-BASED APPLICATION FOR FORECASTING PRICES OF LUMP SUM PAY ITEMS

PURPOSE OF THE TOOL

The purpose of the development of a web application tool was to provide a data-driven tool for estimating the prices of major lump sum items (i.e., Traffic Control 150-1000 and Grading Complete 210-0100) used in highway construction projects in the conceptual stage. Project cost estimating professionals can use this tool to facilitate accurate cost estimation during the early phases of the design development utilizing key information items about the project when design details are not available.

The Lump Sum Item Cost Estimator tool is designed for enhancing the accuracy of early-stage cost estimation for lump sum pay items. After inserting numerical, categorical, and text information of project attributes retrieved from project development documents (i.e., preconstruction status reports, concept reports, and field plan review reports), the users can have instant cost estimation of major lump sum items (i.e., Traffic Control 150-1000 and Grading Complete 210-0100). With the constantly evolving programming landscape, this fully deployed Flask web application was implemented in a Python environment. Despite large model complexity, the tool exhibits high computing speed. This tool aids designers and cost estimators with a flexible and intelligent platform for early cost estimation of two important LS line items, Traffic Control 150-1000 and Grading Complete 210-0100.

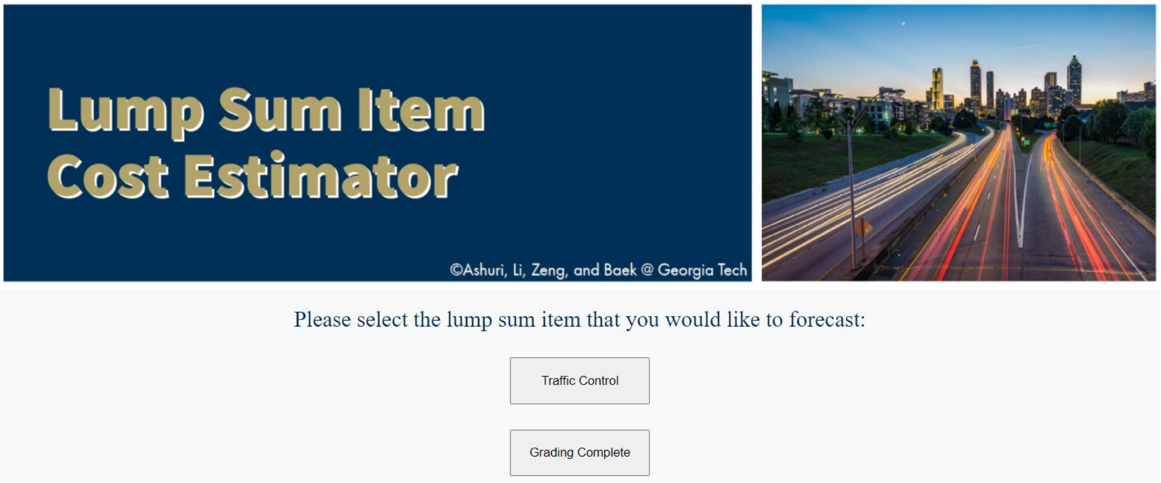
STRUCTURE OF THE TOOL

The tool consists of a *Home* page, *Navigation* page, *Data Inputs* page, and *Results* page, which are introduced below.

Home Page

The *Home* page contains separate links to the *Navigation* pages for Traffic Control (150-1000) and Grading Complete (210-0100) LS items, as shown in figure 11.

- Click the "Traffic Control" button to forecast the price of the Traffic Control (150-1000) lump sum item.
- Click the "Grading Complete" button to forecast the price of the Grading Complete (210-0100) lump sum item.



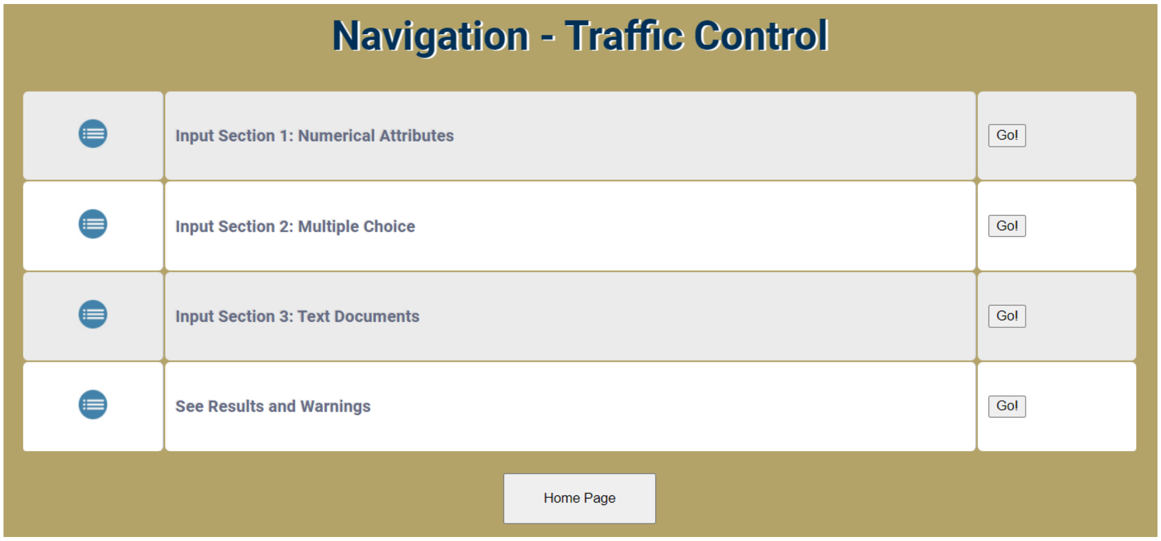
Please select the lump sum item that you would like to forecast:

Figure 11. Screen capture. *Home* page in the Lump Sum Item Cost Estimator.

Navigation Page

After selecting a lump sum item on the *Home* page, the user is directed to the intended *Navigation* page, as shown in Ffigure 12. The *Navigation* page contains links to three different input sections (i.e., numerical attributes, multiple choice, and text documents) and the *Results* page.

- Click the “Go” button on the row of the desired input section to enter the corresponding project attributes. After completion of each input section, use the “Go Back” button to return to the *Navigation* page or the “Continue” button to go to the next input section.
- Click the “Go” button on the row next to the See Results and Warnings option to view the results and any warning information.
- Click the “Home Page” button to return to the *Home* page.



Navigation - Traffic Control

☰	Input Section 1: Numerical Attributes	Go!
☰	Input Section 2: Multiple Choice	Go!
☰	Input Section 3: Text Documents	Go!
☰	See Results and Warnings	Go!

Figure 12. Screen capture. *Navigation* page for Traffic Control in the Lump Sum Item Cost Estimator.

Insert Project Attributes

To insert project attribute, click on the “Go” button next to the desired input type (i.e., numerical, multiple choice, or text) on the *Navigation* page. The data inputs process allows entry of project attributes information through manual-entry fields and drop-down menus. Screenshots of the attributes are provided in figure 13 and figure 14. Fill out or select the required information as indicated in each input section and according to the following detailed instructions of the project attributes to be inserted.[1]

Numerical Attributes

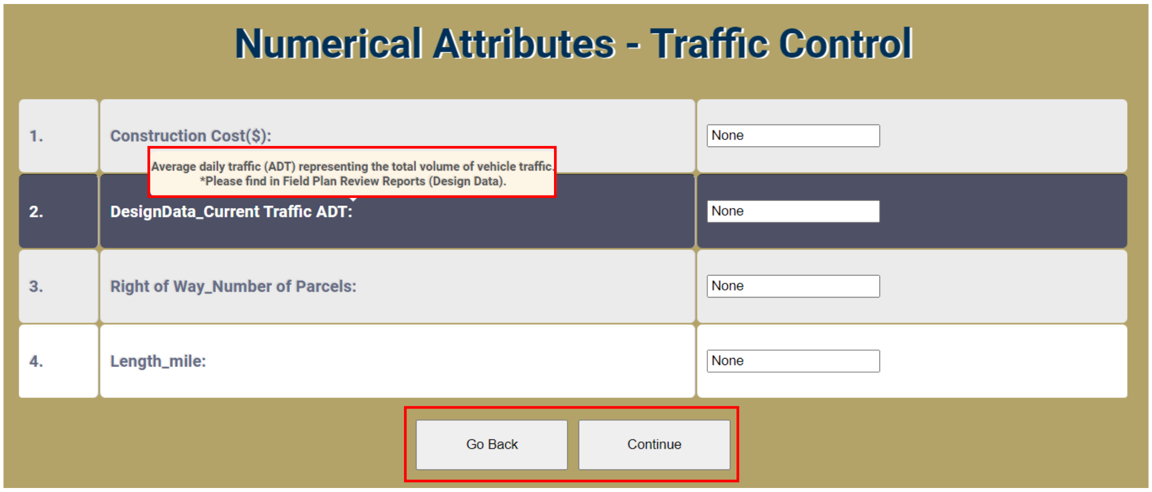
The *Numerical Attributes* page allows entry of numerical attributes, which include (1) Construction Cost ($), (2) DesignData_Current Traffic ADT, (3) Right of Way_Number of Parcels, and (4) Length_mile. These options are available for both Traffic Control (150-1000) and Grading Complete (210-0100) (as selected initially on the *Home* page).

- *Construction Cost ($):* Construction cost includes construction, 5 percent Engineering and Inspection, Contingencies, and Liquid AC Cost Adjustment (not including ROW, Reimbursable Utility, PE, Environmental Mitigation Costs).

[1] The user can also refer to the detailed explanation for each attribute by clicking the attribute name: click once to open the popup message and click again to close the message box. For text inputs, the explanation is provided in the placeholder texts.

Related information can be found in the concept report and preconstruction status report (Coordination, Activities, Responsibilities, and Costs).

- *DesignData_Current Traffic ADT:* Average daily traffic represents the total volume of vehicle traffic. Related information can be found in the Field Plan Review Reports (Design Data).
- *Right of Way_Number of Parcels:* Number of parcels for the right-of-way is required in a project. Related information can be found in the Field Plan Review Reports (Right of Way).
- *Length_mile:* Length of the project represents the total miles of the project. Related information can be found in the Preconstruction Status Report.



Numerical Attributes - Traffic Control

1.	Construction Cost($):	None
2.	DesignData_Current Traffic ADT:	None
3.	Right of Way_Number of Parcels:	None
4.	Length_mile:	None

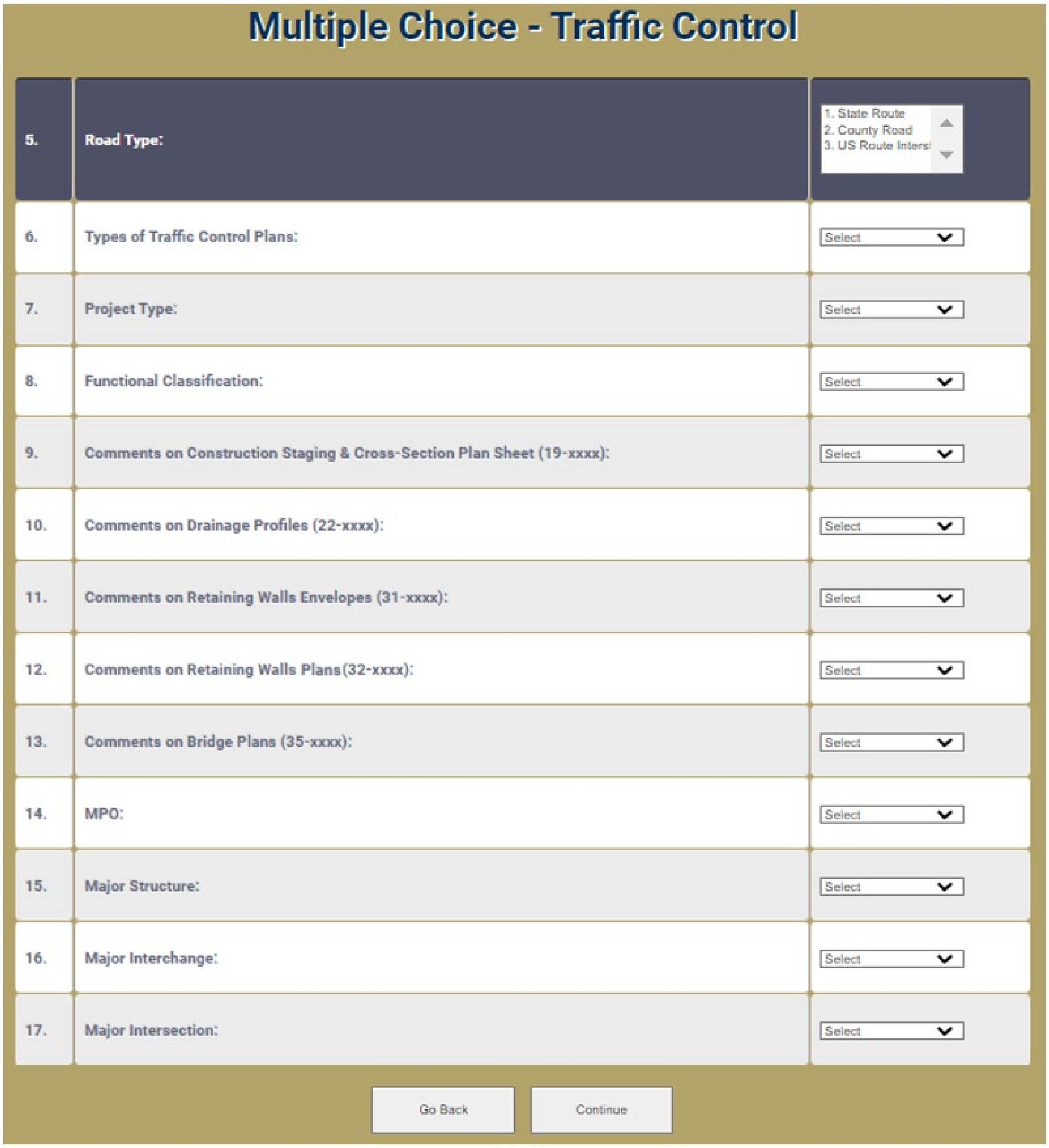
Average daily traffic (ADT) representing the total volume of vehicle traffic.
*Please find in Field Plan Review Reports (Design Data).

Figure 13. Screen capture. *Numerical Attributes* page in the Lump Sum Item Cost Estimator.

Multiple Choice

The *Multiple Choice* page contains 13 attributes for selection from the provided options. These options are available for both Traffic Control (150-1000) and Grading Complete (210-0100) (as selected initially on the *Home* page).

- *Road Type:* Select the types of roadways based on the project description in the Preconstruction Status Reports (PI_Description). To select multiple options, hold down the Control (Ctrl) key.
- *Types of Traffic Control Plans:* Select from several types of traffic control plans. Related information can be found in the Field Plan Review Reports (Special Provisions).
- *Project Type:* Select the project type from the options provided. Related information can be found in the Field Plan Review Reports (Project Description).
- *Functional Classification:* Select the functional classification for the project. Related information can be found in the Field Plan Review Reports.
- *Comments on Construction Staging & Cross-Section Plan Sheet (19-xxxx):* Select options related to the existence of potential issues that affect constructability or construction schedules. Related information can be found in the Concept Reports (Construction).



Multiple Choice - Traffic Control

5.	Road Type:	1. State Route 2. County Road 3. US Route Inters
6.	Types of Traffic Control Plans:	Select
7.	Project Type:	Select
8.	Functional Classification:	Select
9.	Comments on Construction Staging & Cross-Section Plan Sheet (19-xxxx):	Select
10.	Comments on Drainage Profiles (22-xxxx):	Select
11.	Comments on Retaining Walls Envelopes (31-xxxx):	Select
12.	Comments on Retaining Walls Plans(32-xxxx):	Select
13.	Comments on Bridge Plans (35-xxxx):	Select
14.	MPO:	Select
15.	Major Structure:	Select
16.	Major Interchange:	Select
17.	Major Intersection:	Select

Figure 14. Screen capture. *Multiple Choice* page in the Lump Sum Item Cost Estimator.

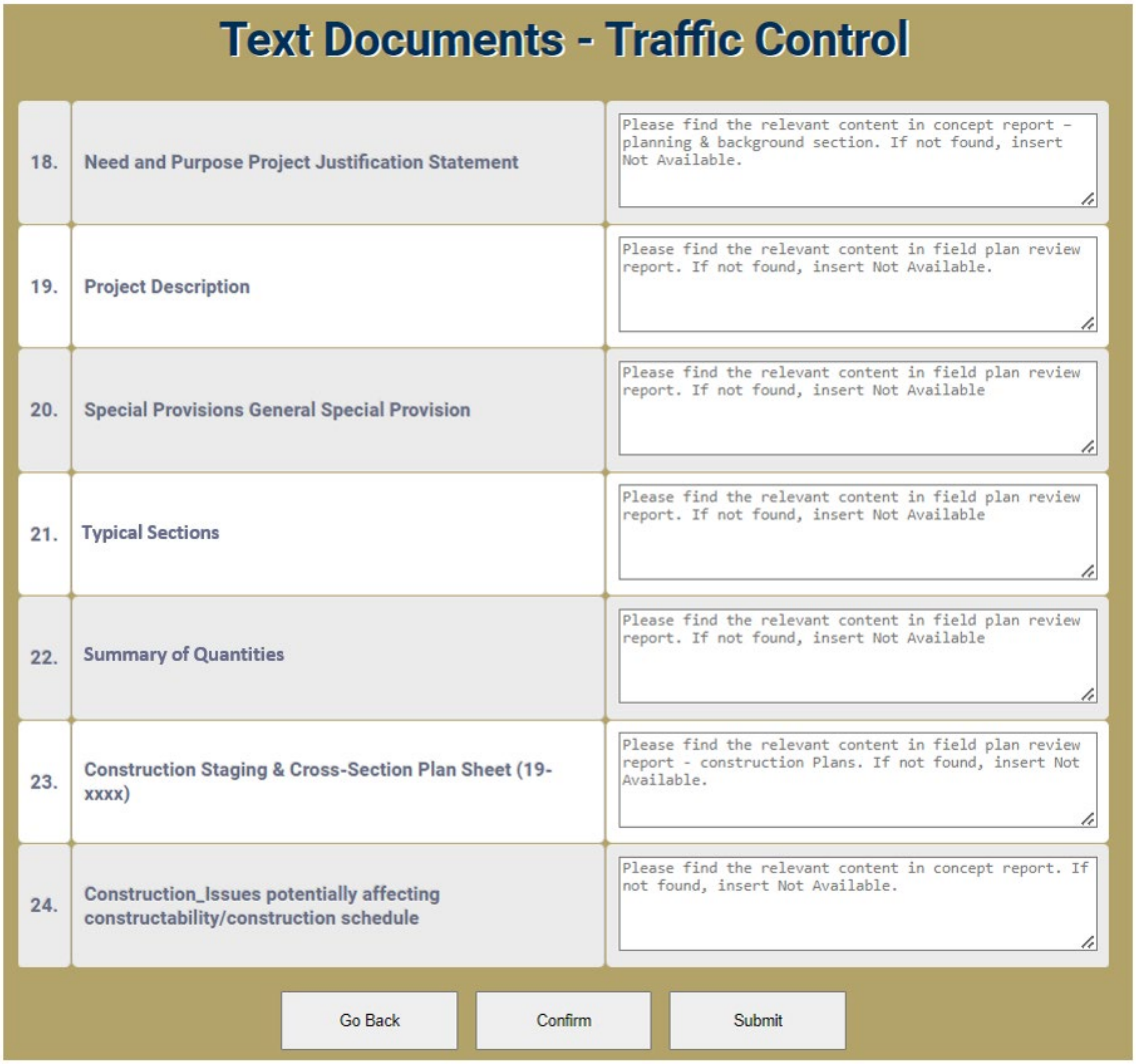
- *Comments on Drainage Profiles (22-xxxx):* Select options for comments related to Drainage Profiles (22-xxxx) in the Field Plan Review Reports. Related information can be found in the Field Plan Review Reports (Construction Plans).

- *Comments on Retaining Walls Envelopes (31-xxxx):* Select options for comments related to Retaining Walls Envelopes (31-xxxx) in the Field Plan Review Reports. Related information can be found in the Field Plan Review Reports (Construction Plans).
- *Comments on Retaining Walls Plans (32-xxxx):* Select options for comments related to Retaining Walls Plans (32-xxxx) in the Field Plan Review Reports. Related information can be found in the Field Plan Review Reports (Construction Plans).
- *Comments on Bridge Plans (35-xxxx):* Select options for comments related to Bridge Plans (35-xxxx) in the Field Plan Review Reports. Related information can be found in the Field Plan Review Reports (Construction Plans).
- *MPO:* Select options regarding areas with a population greater than 50,000, defined by the U.S. Census. Related information can be found in the Preconstruction Status Report.
- *Major Structure:* Select options regarding the existence of major structures, such as bridges and retaining walls. Related information can be found in the Concept Report (Design and Structure Section).
- *Major Interchange:* Select options regarding the existence of major interchanges in the project location. Related information can be found in the Concept Report (Design and Structure Section).

- *Major Intersection:* Select options regarding the existence of major intersections in the project location. Related information can be found in the Concept Report (Design and Structure Section).

Text Documents

The *Text Documents* page contains seven different text attributes for Traffic Control (150-1000) and Grading Complete (210-0100), as shown in **Error! Reference source not found.** and **Error! Reference source not found.**. To insert text inputs, copy the entire paragraph for each text attribute and paste it into the corresponding position; no further edit is needed. If the required information is not found, insert “Not Available”.



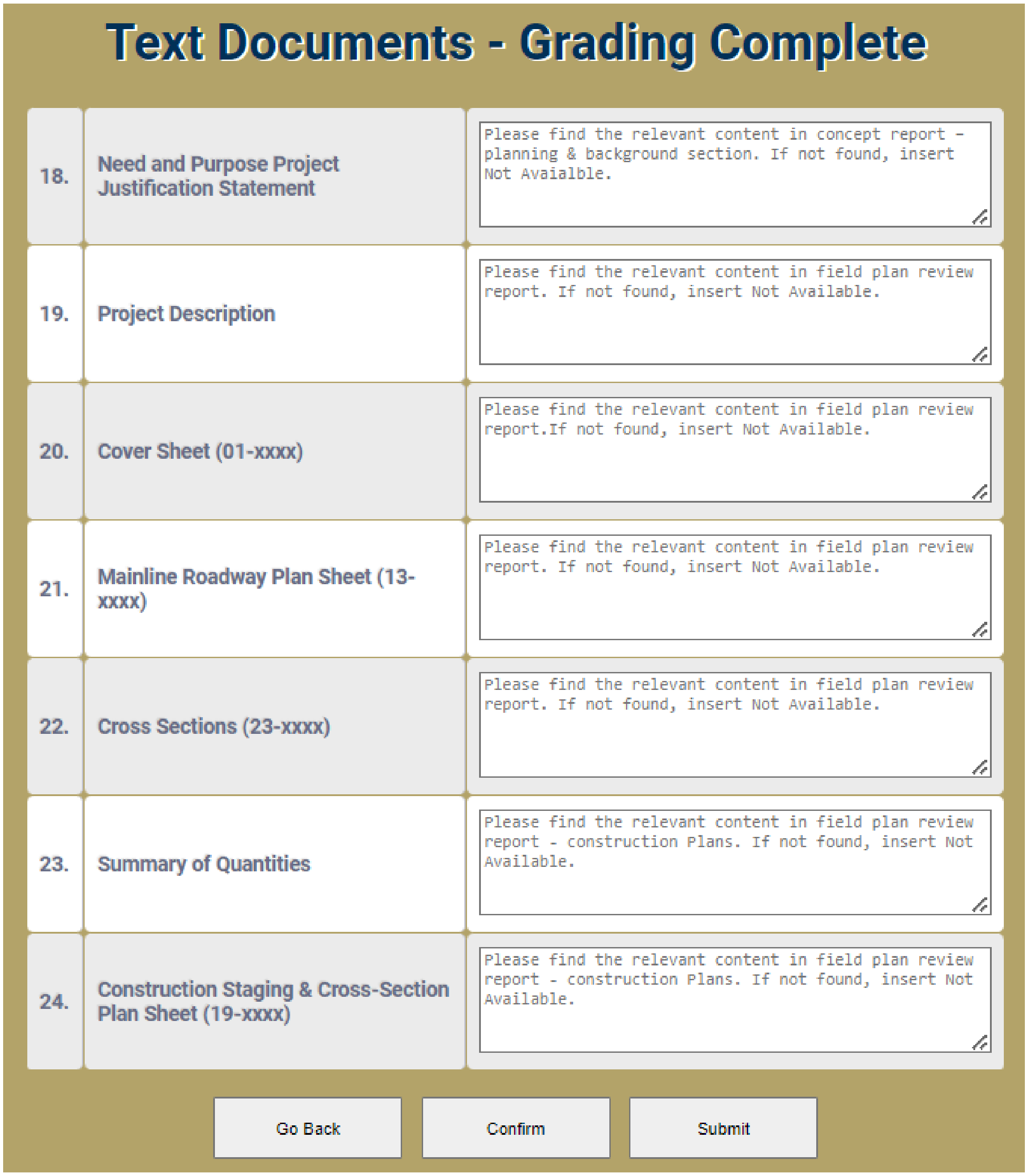
Text Documents - Traffic Control

18.	Need and Purpose Project Justification Statement	Please find the relevant content in concept report - planning & background section. If not found, insert Not Available.
19.	Project Description	Please find the relevant content in field plan review report. If not found, insert Not Available.
20.	Special Provisions General Special Provision	Please find the relevant content in field plan review report. If not found, insert Not Available
21.	Typical Sections	Please find the relevant content in field plan review report. If not found, insert Not Available
22.	Summary of Quantities	Please find the relevant content in field plan review report. If not found, insert Not Available
23.	Construction Staging & Cross-Section Plan Sheet (19-xxxx)	Please find the relevant content in field plan review report - construction Plans. If not found, insert Not Available.
24.	Construction_Issues potentially affecting constructability/construction schedule	Please find the relevant content in concept report. If not found, insert Not Available.

Figure 15. Screen capture. *Text Documents – Traffic Control* page in the Lump Sum Item Cost Estimator.

- *Text Documents – Traffic Control*
  - *Need and Purpose Project Justification Statement:* Related information can be found in the Concept Reports (Planning & Background Section).
  - *Project Description:* Related information can be found in the Field Plan Review Reports.

- *Special Provisions General Special Provision:* Related information can be found in the Field Plan Review Reports.
- *Typical Sections:* Related information can be found in the Field Plan Review Reports.
- *Summary of Quantities:* Related information can be found in the Field Plan Review Reports.
- *Construction Staging & Cross-Section Plan Sheet (19-xxxx):* Related information can be found in the Field Plan Review Reports (Construction Plans).
- *Construction_Issues potentially affecting constructability/construction schedule*: Related information can be found in the Concept Reports.



Text Documents - Grading Complete

18.	Need and Purpose Project Justification Statement	Please find the relevant content in concept report - planning & background section. If not found, insert Not Avaialble.
19.	Project Description	Please find the relevant content in field plan review report. If not found, insert Not Available.
20.	Cover Sheet (01-xxxx)	Please find the relevant content in field plan review report.If not found, insert Not Available.
21.	Mainline Roadway Plan Sheet (13-xxxx)	Please find the relevant content in field plan review report. If not found, insert Not Available.
22.	Cross Sections (23-xxxx)	Please find the relevant content in field plan review report. If not found, insert Not Available.
23.	Summary of Quantities	Please find the relevant content in field plan review report - construction Plans. If not found, insert Not Available.
24.	Construction Staging & Cross-Section Plan Sheet (19-xxxx)	Please find the relevant content in field plan review report - construction Plans. If not found, insert Not Available.

Figure 16. Screen capture. *Text Documents – Grading Complete* page in the Lump Sum Item Cost Estimator.

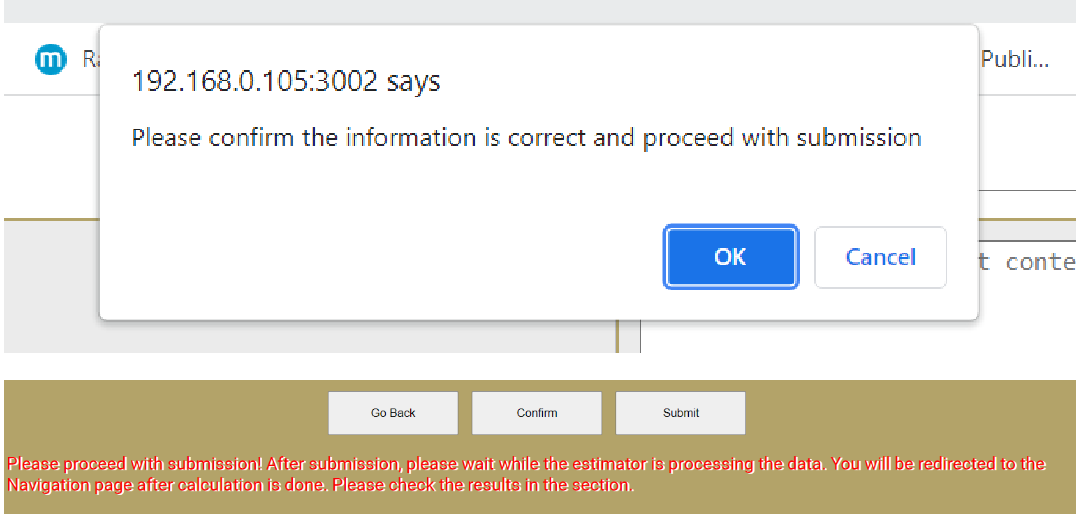
- *Text Documents – Grading Complete*
  - *Need and Purpose Project Justification Statement:* Related information can be found in the Concept Reports (Planning & Background Section).

  - *Project Description:* Related information can be found in the Field Plan Review Reports.
  - *Cover Sheet (01-xxxx):* Related information can be found in the Field Plan Review Reports.
  - *Mainline Roadway Plan Sheet (13-xxxx):* Related information can be found in the Field Plan Review Reports.
  - *Cross Sections (23-xxxx):* Related information can be found in the Field Plan Review Reports.
  - *Summary of Quantities:* Related information can be found in the Field Plan Review Reports.
  - *Construction Staging & Cross-Section Plan Sheet (19-xxxx):* Related information can be found in the Field Plan Review Reports (Construction Plans).

After filling in all the information on the appropriate *Text Documents* page:

- Click the “Confirm” button and the “OK” button in the popup message, as shown in figure 17. The instructions for submission will show up at the bottom of the page.
- Click the “Submit” button to submit the project attributes for the tool to process the data. The computing usually takes a few seconds; afterward, the tool will automatically redirect to the *Navigation* page, where results and warnings can be

accessed by selecting the "Go" button next to the See Results and Warnings option.



192.168.0.105:3002 says

Please confirm the information is correct and proceed with submission

Please proceed with submission! After submission, please wait while the estimator is processing the data. You will be redirected to the Navigation page after calculation is done. Please check the results in the section.

Figure 17. Screen capture. Confirm Inputs dialog box in the Lump Sum Item Cost Estimator.

Results **Page**

The Lump Sum Cost Estimator *Results* page allows the user to check for any warning messages and view the prediction results. To access the results, click the "Go" button next to the See Results and Warnings option on the *Navigation* page, as shown in figure 18.

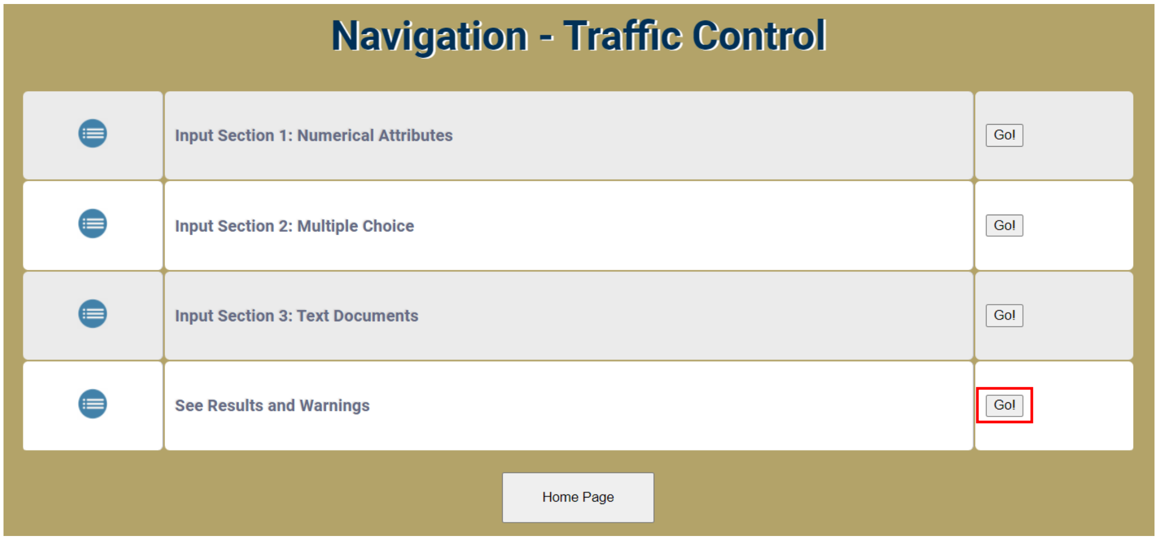
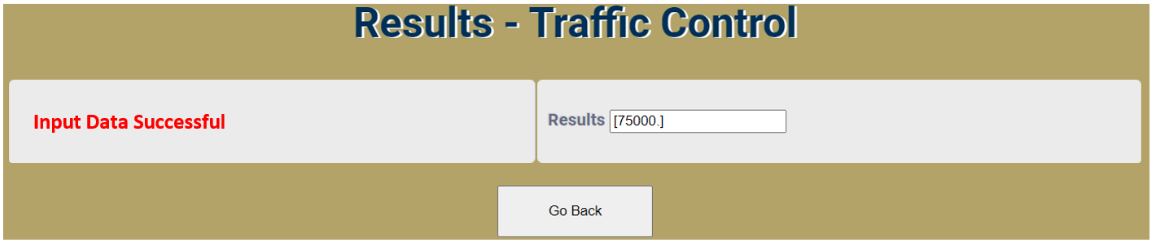

Navigation - Traffic Control
Input Section 1: Numerical Attributes
Go!
Input Section 2: Multiple Choice
Go!
Input Section 3: Text Documents
Go!
See Results and Warnings
Go!
Home Page

Figure 18. Screen capture. See Results and Warnings option in the Lump Sum Item Cost Estimator.

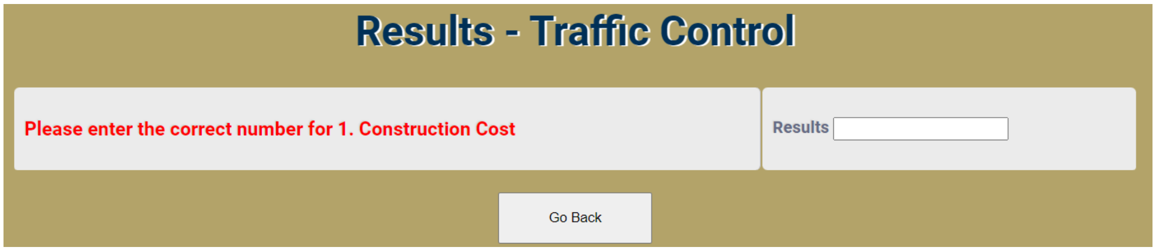
If all required inputs are entered properly, the *Results* page will display the message "Input Data Successful," and the predicted lump sum item price will be provided in the "Results" field, as shown in figure 19. Otherwise, warning information for missing or incorrectly entered items, as shown in figure 20, will be displayed.



Input Data Successful

Results [75000.]

Figure 19. Screen capture. Example results on the *Results* page in the Lump Sum Item Cost Estimator.


Results - Traffic Control

Please enter the correct number for 1. Construction Cost

Figure 20. Screen capture. Example warning on the *Results* page in the Lump Sum Item Cost Estimator.

Start New Calculation

To start a new calculation, on the *Results* page, click the “Go Back” button to return to the *Home* page, then select the lump sum item of interest and repeat the data input process. (Note: The tool will clear the inputs from the last calculation.)

For more information about this tool, contact Dr. Baabak Ashuri at the Georgia Institute of Technology:

Baabak Ashuri, Ph.D., DBIA
Professor
School of Building Construction | School of Civil & Environmental Engineering
Georgia Institute of Technology
Phone: (404) 385-7608
E-mail: baabak@gatech.edu

CHAPTER 7. CONCLUSIONS

The overarching objective of this research project is to develop forecasting models to estimate the prices of the Traffic Control and Grading Complete lump sum pay items using advanced text mining and machine learning algorithms that detect key patterns of information generated during project development and provide higher accuracy of the cost estimates.

To achieve the research objectives, this research used text mining algorithms, including term frequency–inverse document frequency and principal component analysis, to capture key patterns of information from unstructured text files (i.e., concept reports, field plan review reports, and preconstruction status reports). It also used data processing algorithms, including the synthetic minority oversampling technique and the Boruta feature selection algorithm, and machine learning algorithms, including random forest, bagging, k-nearest neighbors, and stacking regressor, to develop forecasting models for the prices of the Traffic Control and Grading Complete LS pay items. This research collected the prices of the Traffic Control and Grading Complete LS pay items used in highway projects in the state of Georgia. With the collected data, a forecasting model for the prices of the Traffic Control and Grading Complete LS pay items was developed.

This research used several machine learning algorithms to develop forecasting models for the segments of the collected data and select the best-performing algorithms for predicting the prices of the Traffic Control and Grading Complete LS pay items for each segment. Based on the mean absolute percentage error, this research found that KNN, random forest, KNN, random forest, and stacking regressor were the best-performing algorithms for predicting the prices of a Traffic Control LS pay item in Segments 1, 2, 3, 4, and 5, respectively. Moreover, this research selected KNN, KNN, and random forest algorithms as the best-performing algorithms in predicting the prices of a Grading Complete LS pay item in Segments 1, 2, and 3, respectively. Next, the accuracy of the forecasting models was compared between partitioned data and data without partitioning. The results of the model comparison indicated that the newly developed machine learning models for forecasting the prices of the Traffic Control and Grading Complete LS pay items in the defined segments showed a higher level of forecasting accuracy.

Finally, a web-based application tool was developed in a Python environment to help designers developing cost estimates with a data-driven tool for estimating the prices of the Traffic Control and Grading Complete LS pay items. This tool serves as a stepping-stone for transforming the traditional methodology of cost estimation, which heavily relies on designers’ experience, into a more flexible and intelligent solution.

ACKNOWLEDGMENTS

The research reported herein was sponsored by the Georgia Department of Transportation through Research Project 20-17. The authors acknowledge and appreciate the help from Mr. Erik Rohde, State Project Review Engineer, and Mr. Sam Woods, Assistant Office Head of Roadway Design. Their invaluable assistance and insight were critical to the completion of this research. Mr. Buchanan Teague, Head of Office of Information Technology (IT); Mr. Prashant Kasurde; and Mr. Ezhilarasan Ethuraj shared their knowledge and experience on the development of a web-based application for lump sum pay items throughout the course of this project. The authors also thank Ms. Supriya Kamatkar, Assistant Office Head (Research), Office of Performance-based Management and Research, and Research Project Managers, Ms. Sarah Lamothe, Mr. Sunil Thapa, and Mr. Bradley Bilsback who have helped us throughout the progress of this research project.

REFERENCES

Adel, K., Elyamany, A., Belal, A.M., and Kotb, A.S. (2016). "Developing Parametric Model for Conceptual Cost Estimate of Highway Projects." *International Journal of Engineering Science and Computing*, *6*(7), pp. 1728–1734.

Al-Tabtabai, H., Alex, A. P., and Tantash, M. (1999). Preliminary cost estimation of highway construction using neural networks. *Cost Engineering, 41(3),* 19.

Anderson, S.D., Molenaar, K.R., and Schexnayder, C.J. (2007). *Guidance for Cost Estimation and Management for Highway Projects During Planning, Programming, and Preconstruction*. NCHRP Report 574, Transportation Research Board, Washington, DC. Available online: https://dx.doi.org/10.17226/14014.

Assaad, R. and El-adaway, I.H. (2020). "Bridge Infrastructure Asset Management System: Comparative Computational Machine Learning Approach for Evaluating and Predicting Deck Deterioration Conditions." *Journal of Infrastructure Systems*, *26*(3), 04020032. Available online: http://dx.doi.org/10.1061/(ASCE)IS.1943-555X.0000572.

Baek, M., Mostaan, K., and Ashuri, B. (2016). "Recommended Practices for the Cost Control of Highway Project Development." *Proceedings* of the Construction Research Congress 2016, San Juan, Puerto Rico, May 31–June 2, pp. 739–748. Available online: http://dx.doi.org/10.1061/9780784479827.075.

Baek, M., and Ashuri, B. (2018). Assessment of spatial correlation patterns of unit price bids and external factors. In Proc., 54th Associated Schools of Construction (ASC) Annual Int. Conf (pp. 18-21).

Baek, M., and Ashuri, B. (2019). Analysis of the variability of submitted unit price bids for asphalt line items in highway projects. Journal of Construction Engineering and Management, 145(4), 04019020.

Baek, M. and Ashuri, B. (2021). "Synthesis of Practices and Tools for Cost Estimation and Cost Management for Transportation Projects." *Journal for the Advancement of Performance Information and Value*, *13*(1), 22-22.

Blampied, N.B. (2018). *Parametric Functions for Conceptual and Feasibility Estimating in Public Highway Project Portfolios*. University of California, Berkeley.

Breiman, L. (1996). *Bagging predictors*. Machine learning, 24(2), 123-140.

Cao, Y., and Ashuri, B. (2020). "Predicting the volatility of highway construction cost index using long short-term memory." *Journal of Management in Engineering, 36(4),* 04020020.

Cao, Y., Ashuri, B., and Baek, M. (2018). "Prediction of Unit Price Bids of Resurfacing Highway Projects Through Ensemble Machine Learning." *Journal of Computing in Civil Engineering*, *32*(5), 04018043. Available online: http://dx.doi.org/10.1061/(ASCE)CP.1943-5487.0000788.

Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. (2002). "SMOTE: Synthetic Minority Over-sampling Technique." *Journal of Artificial Intelligence Research*, *16*, pp. 321–357. Available online: https://doi.org/10.1613/jair.953.

Cheng, M. Y., Tsai, H. C., and Sudjono, E. (2010). Conceptual cost estimates using evolutionary fuzzy hybrid neural network for projects in construction industry. *Expert Systems with Applications, 37(6),* 4224-4231.

Chou, J.-S. (2009). "Web-based CBR System Applied to Early Cost Budgeting for Pavement Maintenance Project." *Expert Systems with Applications*, *36*(2), pp. 2947–2960. Available online: https://doi.org/10.1016/j.eswa.2008.01.025.

Chou, J.-S., Peng, M., Persad, K.R., and O'Connor, J.T. (2006). "Quantity-based Approach to Preliminary Cost Estimates for Highway Projects." *Transportation Research Record: Journal of the Transportation Research Board*, *1946*(1), pp. 22–30. Available online: https://doi.org/10.1177%2F0361198106194600103.

Creese, R.C. and Li, L. (1995). "Cost Estimation of Timber Bridges Using Neural Networks." *Cost Engineering*, *37*(5), pp. 17–22.

De la Garza, J.M. and Rouhana, K.G. (1995). "Neural Networks Versus Parameter-based Applications in Cost Estimating." *Cost Engineering*, *37*(2), pp. 14–18.

Elmousalami, H.H. (2020). "Artificial Intelligence and Parametric Construction Cost Estimate Modeling: State-of-the-Art Review." *Journal of Construction Engineering and Management*, *146*(1), 03119008. Available online: https://doi.org/10.1061/(ASCE)CO.1943-7862.0001678.

Gardner, B. J. (2015). "Applying artificial neural networks to top-down construction cost estimating of highway projects at the conceptual stage." *Iowa State University*.

Gardner, B. J., Gransberg, D. D., and Jeong, H. D. (2016). Reducing data-collection efforts for conceptual cost estimating

Guyon, I. and Elisseeff, A. (2003). "An Introduction to Variable and Feature Selection." *Journal of Machine Learning Research*, *3*(Mar), pp. 1157–1182. Available online: https://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf.

Hegazy, T. and Ayed, A. (1998). "Neural Network Model for Parametric Cost Estimation of Highway Projects." *Journal of Construction Engineering and Management*, *124*(3), pp. 210–218.

Jafari, P., Al Hattab, M., Mohamed, E., and AbouRizk, S. (2021). "Automated Extraction and Time-Cost Prediction of Contractual Reporting Requirements in Construction Using Natural Language Processing and Simulation." *Applied Sciences*, *11*(13), 6188. Available online: https://doi.org/10.3390/app11136188.

Kim, H.J., Seo, Y.C., and Hyun, C.T. (2012). "A Hybrid Conceptual Cost Estimating Model for Large Building Projects." *Automation in Construction*, *25*, pp. 72–81. Available online: https://doi.org/10.1016/J.AUTCON.2012.04.006.

Kohavi, R. and John, G.H. (1997). "Wrappers for Feature Subset Selection." *Artificial Intelligence*, *97*(1–2), pp. 273–324. Available online: https://doi.org/10.1016/S0004-3702(97)00043-X.

Kotekar, S. and Kamath, S.S. (2018). "Enhancing Web Service Discovery Using Meta-heuristic CSO and PCA Based Clustering." In *Progress in Intelligent Computing Techniques: Theory, Practice, and Applications*, Springer, Singapore, pp. 393–403.

Kursa, M.B. and Rudnicki, W.R. (2010). “Feature Selection with the Boruta Package.” *Journal of Statistical Software*, *36*(11), pp. 1–13. Available online: https://doi.org/10.18637/jss.v036.i11.

Li, M., and Ashuri, B. (2021). “Proportional Cox Hazards Model to Quantify the Likelihood of Underestimation in Transportation Projects.” *Journal of Construction Engineering and Management, 147(10),* 04021134.

Li, M., M. Baek, and B. Ashuri. "Forecasting ratio of low bid to owner’s estimate for highway construction." *Journal of Construction Engineering and Management* 147, no. 1 (2021): 04020157.

Lin, C.T., Wang, N.J., Xiao, H., and Eckert, C. (2015). “Feature Selection and Extraction for Malware Classification.” *Journal of Information Science and Engineering*, *31*(3), pp. 965–992.

Mahamid, I. (2011). “Early Cost Estimating for Road Construction Projects Using Multiple Regression Techniques.” *Australasian Journal of Construction Economics and Building*, *11*(4), pp. 87–101.

Mahamid, I. (2013). “Conceptual Cost Estimate of Road Construction Projects in Saudi Arabia.” *Jordan Journal of Civil Engineering*, *7*(3), pp. 285–294.

Mohammadi, S. (2020). “A Test of Harmful Multicollinearity: A Generalized Ridge Regression Approach.” *Communications in Statistics – Theory and Methods*, *51*(3), pp. 1–20. Available online: https://doi.org/10.1080/03610926.2020.1754855.

Moon, S., Lee, G., and Chi, S. (2021). “Semantic Text-pairing for Relevant Provision Identification in Construction Specification Reviews.” *Automation in Construction*, *128*, 103780. Available online: https://doi.org/10.1016/j.autcon.2021.103780.

Murphy, K.P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press, Cambridge, MA.

Nasiriany, S., Thomas, G., Wang, W., Yang, A., Listgarten, J., and Sahai, A. (2019). *A Comprehensive Guide to Machine Learning*. University of California at Berkeley, 82–88. Available online: https://snasiriany.me/files/ml-book.pdf.

Ogungbile, A.J., Oke, A.E., and Rasak, K. (2018). “Developing Cost Model for Preliminary Estimate of Road Projects in Nigeria.” *International Journal of Sustainable Real Estate and Construction Economics*, *1*(2), pp. 182–199.

Paulsen, C., Gallivan, F., Chavez, M., and Venner, M. (2008). *NCHRP 8-36 Task 72: Guidelines for Cost Estimation Improvements at State DOTs*. National Cooperative Highway Research Program, Washington, DC. Available online: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.183.2432&rep=rep1&type=pdf.

Petroutsatou, K., Georgopoulos, E., Lambropoulos, S., and Pantouvakis, J.P. (2012). “Early Cost Estimating of Road Tunnel Construction Using Neural Networks.” *Journal of Construction Engineering and Management*, *138*(6), pp. 679–687. Available online: https://doi.org/10.1061/(ASCE)CO.1943-7862.0000479.

Ramsingh, J. and Bhuvaneswari, V. (2021). "An Efficient Map Reduce-based Hybrid NBC-TFIDF Algorithm to Mine the Public Sentiment on Diabetes Mellitus – A Big Data Approach." *Journal of King Saud University – Computer and Information Sciences*, *33*(8), pp. 1018–1029. Available online: https://doi.org/10.1016/j.jksuci.2018.06.011.

Rong, T., Gong, H., and Ng, W.W.Y. (2014). "Stochastic Sensitivity Oversampling Technique for Imbalanced Data." In Wang, X., Pedrycz, W., Chan, P., He, Q. (eds), *Machine Learning and Cybernetics*, ICMLC 2014, *Communications in Computer Science and Information Science*, *481*, pp. 161–171, Springer, Berlin, Heidelberg. Available online: https://doi.org/10.1007/978-3-662-45652-1_18.

Sodikov, J. (2005). "Cost Estimation of Highway Projects in Developing Countries: Artificial Neural Network Approach." *Journal of the Eastern Asia Society for Transportation Studies*, *6*, pp. 1036–1047. Available online: https://doi.org/10.11175/EASTS.6.1036.

Song, Y., Liang, J., Lu, J., and Zhao, X. (2017). "An Efficient Instance Selection Algorithm for k Nearest Neighbor Regression." *Neurocomputing*, *251*, pp. 26–34. Available online: https://doi.org/10.1016/j.neucom.2017.04.018.

Su, C.T. and Hsiao, Y.H. (2007). "An Evaluation of the Robustness of MTS for Imbalanced Data." *IEEE Transactions on Knowledge and Data Engineering*, *19*(10), pp. 1321–1332. Available online: https://doi.org/10.1109/TKDE.2007.190623.

Tijanić, K., Car-Pušić, D., and Šperac, M. (2019). "Cost Estimation in Road Construction Using Artificial Neural Network." *Neural Computing and Applications*, pp. 1–13. Available online: https://doi.org/10.1007/s00521-019-04443-y.

Trost, S.M. and Oberlender, G.D. (2003). "Predicting Accuracy of Early Cost Estimates Using Factor Analysis and Multivariate Regression." *Journal of Construction Engineering and Management*, *129*(2), pp. 198–204. Available online: https://doi.org/10.1061/(ASCE)0733-9364(2003)129:2(198).

ul Hassan, F., Le, T., and Tran, D.-H. (2020). "Multi-class Categorization of Design–Build Contract Requirements Using Text Mining and Natural Language Processing Techniques." *Construction Research Congress 2020: Project Management and Controls, Materials, and Contracts*, pp. 1266–1274, American Society of Civil Engineers, Reston, VA. Available online: https://doi.org/10.1061/9780784482889.135.