Data-Driven Predictive Modeling of Highway Construction Cost Items

Article *in* Journal of Construction Engineering and Management · December 2020

DOI: 10.1061/(ASCE)CO.1943-7862.0001991

5 authors, including:

Amirsaman Mahdavian
University of Central Florida
15 PUBLICATIONS **29** CITATIONS
SEE PROFILE

Alireza Shojaei
Virginia Polytechnic Institute and State University
40 PUBLICATIONS **140** CITATIONS
SEE PROFILE

Milad Salem
University of Central Florida
13 PUBLICATIONS **155** CITATIONS
SEE PROFILE

Data-Driven Predictive Modeling of Highway Construction Cost Items

Amirsaman Mahdavian, A.M.ASCE[1]; Alireza Shojaei, A.M.ASCE[2]; Milad Salem[3]; Jiann Shiun Yuan[4]; and Amr A. Oloufa, M.ASCE[5]

Abstract: The highway network is an economically necessary form of transportation that has a significant impact on the quality of the life of the citizens who use it. Cost overruns in highway projects have been a universal occurrence that jeopardize the development, maintenance, and expansion of this vital infrastructure. Incorrect cost estimations can drive decision makers to pass ineffective policies that have played a large role in the cost overruns of transportation construction projects. The existing prediction models in the literature are limited in one or multiple areas of modeling approach, inputs, and model development robustness. In this research, a model was developed to accurately predict the total construction cost of highway projects by utilizing machine learning algorithms. This study developed a modeling pipeline to automate much of the cost forecasting process, reducing the amount of manual work and dependence on skilled data scientists. This study used the Florida Department of Transportation's (FDOT's) critical highway construction cost items between 2001 and 2017 to test the model. The highways of Florida were selected for testing due to the states' population growth, high immigrant population, logistics, and hurricane frequency. This study used a pool of five categories of independent variables (69 variables total), including the construction market, energy market, socioeconomics, US economy, and temporal variables, which were compiled from relevant sources and existing literature. The results revealed that our linear model exhibits superiority in generalization and prediction of cost items over nonlinear models and is capable of accurately forecasting highway construction costs. Our suggested approach in this study also provides more accurate forecasts for the detailed cost estimation by considering the monthly historical information for the average 92.6% of the six highway construction types mentioned with a 92.51% prediction accuracy. By employing our developed model, local governments, network operators, contractors, and logistics sectors would be capable of a more exact prediction of highway construction costs. **DOI: 10.1061/(ASCE)CO.1943-7862.0001991.**

Author keywords: Data-driven predictive modeling; Construction cost; Highway construction; Cost forecast; Regression analysis; Machine learning.

Introduction

Civil infrastructures are the combination of systems and citizens. Of the various types of infrastructures, the US Bureau of Economic Analysis estimated the total value of US roads at approximately $2.6 trillion. There is no debate as to the necessity of infrastructure financing in the United States and city administrators continue to investigate infrastructure upgrades. As a result, the demand for higher infrastructure spending is one of the few problems that unites both main political bodies in the US (PWC 2016). However, the federal government faces two obstacles in improving the highway network, namely radical investment gaps to construct required highways, and cost overruns of current projects. ASCE awarded the United States road infrastructure a D+ in 2017 and forecasted extreme financing gaps in transportation network development. The anticipated highway infrastructure investment gaps are expected to impose significant losses on the US economy with an approximately $4 trillion loss in gross domestic product (GDP) by 2025, and an approximately $18 trillion loss in GDP over 25 years from 2016 to 2040. Cities worldwide are increasingly eager to perform actions to address these difficulties; however, insufficient funds are a restraining factor and only 16% of cities can self-fund the required infrastructure projects (Kellerman 2018).

Project costs are defined as economic resources in the form of capital, land, labor, and materials that are required for administering, operating, maintaining, improving, enhancing, and expanding the region regarding proposed projects over their life cycle. Highway construction costs can be forecasted by utilizing historical cost items data. A predicted cost item is a valuable tool for planning the required highway construction projects more efficiently. However, the level of complexity of highway cost patterns leads to the need to reconsider highway cost prediction issues using deep structure models with a higher quantity of cost data that consider a greater number of independent variables compared to previous studies.

Cost overruns on highway projects have been a universal occurrence that trouble federal and state organizations worldwide (Flyvbjerg et al. 2002). In such an environment, investors need to consider different risks and high volatilities, such as currency

[1]Ph.D. Candidate, Dept. of Civil Engineering, College of Engineering and Computer Science, Univ. of Central Florida, Orlando, FL 32816 (corresponding author). ORCID: https://orcid.org/0000-0003-2146-4405. Email: amirsaman@knights.ucf.edu

[2]Dept. of Building Construction Science, College of Architecture, Art and Design, Mississippi State Univ., P.O. Box 6222, Mississippi State, MS 39762. Email: shojaei@caad.msstate.edu

[3]Ph.D. Candidate, Dept. of Electrical and Computer Engineering, Univ. of Central Florida, Orlando, FL 32816. ORCID: https://orcid.org/0000-0002-6703-6839. Email: miladsalem@knights.ucf.edu

[4]Professor, Dept. of Electrical and Computer Engineering, Univ. of Central Florida, Orlando, FL 32816. Email: jiann-shiun.yuan@ucf.edu

[5]Professor, Dept. of Engineering and Computer Science, Univ. of Central Florida, Orlando, FL 32816. Email: amr.oloufa@ucf.edu

Note. This manuscript was submitted on April 29, 2020; approved on September 15, 2020; published online on December 24, 2020. Discussion period open until May 24, 2021; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Construction Engineering and Management***, © ASCE, ISSN 0733-9364.

fluctuations that are difficult to forecast (Kaliba et al. 2009). Contractors also must use competing bidding strategies to offer the lowest bid and acquire contracts in the US (Zhang et al. 2017). As a result, arriving at an accurate and competitive price becomes a necessary but challenging task. In practice, due to numerous unpredictable and disruptive trends, long-term forecasts can be unreliable for practical use. However, if implemented correctly, they can achieve an accuracy level beneficial to multiple applications. To predict the cost of highway infrastructure expansion, including adding lanes to an existing road or building a new road, a prediction model is required that considers the construction cost, road maintenance costs, and operating ownership costs. This study aims to predict the future highway construction costs by using internal and external variables and by developing and comparing a variety of nonlinear and linear models that can forecast the cost with high accuracy. In this process, a pipeline containing feature selection was created and optimized to assist the training of the models.

Literature Review

Designing, building, operating, and maintaining a facility requires a reasonable understanding of the anticipated revenues and costs required to form important decisions (Victoria Transport Policy Institute 2016). Turochy et al. (2001) studied the methods of US DOTs in predicting roadway construction budget plans during the initial stages of the project and classified them into two classes. First, multiple DOTs utilize cost-per-unit, which occasionally demands engineering expertise and knowledge to adapt the cost items from the charts. Second, some DOTs use a rough order of estimations of the number of significant payments employing the historical information of similar size projects. A statewide standard method for cost estimation lacks in several DOTs. Those DOTs employ various methods (including those mentioned previously) or techniques that are entirely reliant on engineering practice (Zhang et al. 2017).

Regression models (Wilmot and Cheng 2003) and artificial neural networks (ANNs) (Wilmot and Mei 2005; Shojaei and Mahdavian 2019) are two broadly employed techniques for predicting construction costs. Emsley et al. (2002) also employed linear regression and ANNs to predict construction costs. The results revealed that a significant benefit of the neural network (NN) method was their ability to model the nonlinearity in the data.

Concerning the accuracy, Membah and Asa (2015) determined that the unit cost method was not reliable and led to significant budget underestimations or failed to properly reflect the risks in roadway projects. Therefore, companies often include a separate risk analysis in their estimates before submitting a bid. The unit cost method is also incapable of handling large quantities of data. Schach and Naumann (2007) have claimed that probabilistic methods require skilled workers and data with sufficient quantity and quality. Swei et al. (2017) also developed a parametric method for predicting construction costs that combined a least angle regression for dimensionality reduction and a maximum likelihood estimator for data transformations and tested the framework on 15 different pavement bid items.

A literature review revealed that in total, 20 predictor variables were used in previous studies to investigate the highway construction cost. These included average hourly earnings, average weekly hours, employment level in construction, Dow Jones industrial average, crude oil price, federal funds rate, construction spending, money supply, GDP implicit price deflator, gross domestic product, consumer price index, unemployment rate, building permit number, number of housing starts, prime loan rate, number of bidders, contract duration, producer price index, contractor influence, and months of the year. As a result, based on the identified categories and related variables, this study strived to collect even more information (69 predictor variables) to investigate the subject in a more comprehensive way. Appendix I shows a comprehensive list of the predictor variables used.

Minchin et al. (2004) utilized empirical data from state DOTs and generated a regression model. The number of bidders was recognized as the most crucial parameter in affecting the difference between the lowest bid and the engineer's prediction. Mahamid (2011) employed a multivariable regression model to predict the initial costs of road construction. It was observed that the model utilizing bid quantity as an independent factor achieved superior results compared to models using road length and width. Various studies have examined the construction cost behavior against construction market and macroeconomic independent predictors and the leading parameters have a high correlation to the price of overhead and resources. The economy is widely considered to have a significant impact on highway construction costs (Anderson et al. 2007; Williams 2003), and the bulk of roadway construction costs incorporates human resources, machineries, and materials, which are strictly correlated with the status of the economy (Zhang et al. 2017). Williams (1994) predicted changes in the *Engineering News-Record* construction cost index by utilizing trends in housing starts, prime lending rate, and months of the year as inputs to backpropagation network models.

Wong and Ng (2010) concluded that the construction market and macroeconomic parameters were essential for predicting tender price index. Shahandashti and Ashuri (2016) employed 16 variables (more than any other study examined) and determined that average hourly earnings and crude oil prices are the principal factors affecting highway construction costs. To identify the suitable features for modeling construction cost, some studies (Lowe et al. 2006; Ji and Park 2010; Kim and Hong 2012) used stepwise multiple linear regression (MLR) to make predictions. Shahandashti and Ashuri (2016) also employed vector error correction models to forecast the leading factors affecting highway construction costs.

A review of the literature also demonstrated there had not been much research thus far on the development of a universal automated framework for highway construction cost prediction that incorporates a broad dataset (Florida highways between 2001 and 2017) and inclusive predictors (69 independent variables) utilizing both the linear (five algorithms) and nonlinear (four algorithms) algorithms employing a robust cross-validation method. The pipeline of this study incorporated a hyperparameter optimization framework (or grid search) to identify the best feature selection method and the modeling approach in order to decrease the mean absolute percentage error (MAPE). The need for such a universal automated framework is evident in the literature as there is an apparent inconsistency in the previous successful approaches in terms of algorithm, feature selection method, and other elements of the cost forecasting pipeline. In other words, depending on the characteristics of the investigated case in each previous study, different algorithms and different final parameters have been found to be the optimal choice. As a result, it can be argued that instead of focusing on optimizing a model for a specific case study, a universal automated framework needs to be developed that can create a customized model based on the specific case under investigation.

This study aims to build a model to fill these identified gaps to help contractors and planners enhance the cost estimation of the projects. The results of this study demonstrate the high accuracy of the developed framework that could be easily generalized to be employed by other users. By following the step by step methodology described in this research, and utilizing data related to their

local predictors and projects users can optimize the highway construction cost prediction model accordingly. The final model and leading factors may vary from the ones selected for the tested predictors and dataset optimized for the state of Florida in this study.

Methodology

This project aims to predict future highway construction costs accurately using internal and external variables. Particularly, this study strives to address the following objectives:

- Evaluate the prediction accuracy of multiple machine learning algorithms considering multiple linear and nonlinear relationships between variables to forecast the cost items;
- Evaluate the influence of temporal predictors on the prediction model on road construction cost items; and
- Investigate the impact of the socioeconomic, energy market, US economy, and construction market on highway construction cost items.

A model was developed to automate the process of training, testing, and feature selection (discussed in the next section). To test this model, our work used the Florida Department of Transportation's (FDOT's) critical highway construction cost items between 2001 and 2017. The highways of Florida were selected for testing due to the states' population growth, immigrant population, logistics, critical locations, and hurricane frequency. The 60 cost items' unit prices of the FDOT historical monthly data used for four- and six-lane urban and rural interstate highway construction and widening scenarios were utilized to test the model.

Fig. 1 shows 60 cost items and their associated cost margin in each type of project. From the 60 cost items (dependent variables) covering 100% of the total cost of six highway expansion types (constructing and widening), 10 cost items' monthly historical data were not available (about 7.4% of the total cost), so the remaining 50 cost items were considered to be fed to the pipeline of the study. These cost items covered about 92.6% of the average total cost of highway construction (both new construction and widening construction projects). The data includes about 17,121 projects in small, medium, and large sizes. Then, our team organized the data, analyzed the features, and prepared a monthly unit price between 2001 to 2017. The dataset included a list of 60 cost items, which covers more than 92% of the scope of cost estimation process for six types of highway construction. In total, this study employed 1,027,260 data points for the 17,121 projects in Florida over 17 years (204 months). Costs were analyzed on a

	Dependent Variables	Constructing New Urban 6L	Constructing New Urban 4L	Constructing New Rural 6L	Constructing New Rural 4L	Widening 6L to 8L	Widening 4L to 6L
1	MAINTENANCE OF TRAFFIC	8.18%	8.18%	4.29%	4.29%	8.17%	8.16%
2	MOBILIZATION	9.00%	9.00%	9.00%	9.00%	8.99%	8.98%
3	SEDIMENT BARRIER	0.16%	0.17%	0.30%	0.43%	0.38%	0.43%
4	FLOATING TURBIDITY BARRIER	0.02%	0.02%	0.04%	0.06%	0.02%	0.03%
5	STAKED TURBIDITY BARRIER-NYL REINF PVC	0.01%	0.01%	0.01%	0.02%	0.01%	0.01%
6	SOIL TRACKING PREVENTION DEVICE	0.02%	0.02%	0.04%	0.05%	0.05%	0.06%
7	LITTER REMOVAL	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
8	MOWING	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
9	CLEARING & GRUBBING	2.53%	2.74%	7.13%	8.30%	2.22%	2.49%
10	REGULAR EXCAVATION	1.61%	1.74%	3.87%	4.41%	2.12%	2.39%
11	BORROW EXCAVATION, TRUCK MEASURE	×	×	×	×	10.76%	12.10%
12	EMBANKMENT	5.11%	4.57%	14.15%	17.25%	2.46%	2.76%
13	TYPE B STABILIZATION	2.03%	1.73%	4.19%	3.58%	2.90%	3.26%
14	OPTIONAL BASE,BASE GROUP 04	2.56%	2.78%	4.94%	4.08%	3.54%	3.98%
15	OPTIONAL BASE,BASE GROUP 09	5.95%	4.33%	11.46%	8.97%	5.52%	6.21%
16	MILLING EXIST ASPH PAVT, 1" AVG DEPTH	×	×	×	×	0.66%	0.74%
17	MILLING EXIST ASPH PAVT, 4" AVG DEPTH	×	×	×	×	3.70%	2.77%
18	SUPERPAVE ASPH CONC, TRAF C,PG76-22,PMA	2.08%	2.25%	22.02%	18.29%	4.30%	4.84%
19	SUPERPAVE ASPH CONC, TRAF D,PG76-22,PMA	9.45%	6.82%	3.70%	2.85%	26.94%	23.45%
20	ASPH CONC FC,INC BIT,FC-5,PG76-22,PMA	1.92%	1.41%	×	0.08%	6.92%	5.86%
21	ASPH CONC FC,TRAFFIC B,FC-9.5,PG 76-22	×	×	×	×	0.07%	0.08%
22	CONC CLASS II, ENDWALLS	1.71%	1.55%	1.05%	1.23%	1.13%	1.27%
23	INLETS, DT BOT, TYPE D	0.06%	0.07%	0.06%	0.07%	0.09%	0.10%
24	INLETS, DT BOT, TYPE E	0.20%	0.22%	0.39%	0.59%	×	×
25	INLETS, BARRIER WALL	2.33%	2.10%	×	×	×	×
26	MANHOLES, J-7	0.11%	0.12%	0.22%	0.26%	0.15%	0.17%
27	DESILTING PIPE, 0 - 24"	×	×	×	×		
28	DESILTING PIPE, 25 - 36"	×	×	×	×		
29	PIPE CULV, OPT MATL, ROUND,24"SD	0.70%	0.59%	1.07%	1.34%	1.53%	1.72%
30	PIPE CULV, OPT MATL, ROUND, 24"S/CD	×	X	0.43%	0.50%	×	×
31	PIPE CULV, OPT MATL, ROUND, 30"S/CD	4.94%	×	×	×	0.21%	0.23%
32	PIPE CULV, OPT MATL, ROUND, 36"S/CD	0%	6.19%	0.62%	0.67%	×	×
33	PIPE CULV, OPT MATL, ROUND, 42"S/CD	7.10%	8.49%	0.14%	0.15%	0.20%	0.23%
34	PIPE CULV, OPT MATL, ROUND, 54"S/CD	0.86%	0.93%	1.66%	1.90%	1.19%	1.34%
35	MITERED END SECT, OPTIONAL RD, 24" SD	0.52%	×	1%	1.17%	1.44%	1.62%
36	MEDIAN CONC BARRIER WALL	5.54%	6%	×	×	×	×
37	SHLDR CONC BARRIER, RIGID-SHLDR	16.83%	18.24%	×	×	×	×
38	CONCRETE DITCH PAVT, NR, 3"	×	×	1.57%	1.46%	×	×
39	RUMBLE STRIPS, GROUND-IN, 16" MIN. WIDTH	0.03%	0.03%	0.06%	0.07%	0.08%	0.09%
40	FENCING, TYPE B, 5.1-6.0', STANDARD	1.47%	1.59%	2.72%	3.85%	0.19%	0.21%
41	FENCE GATE,TYP B,SLIDE/CANT,18.1-20'OPEN						
42	PERFORMANCE TURF	×	×	0.01%	0.02%	×	×
43	PERFORMANCE TURF, SOD	0.41%	0.01%	1.18%	1.48%	0.87%	0.97%
44	SINGLE POST SIGN, F&I GM	0.01%	0.23%	0.01%	×	0.02%	0.02%
45	SINGLE POST SIGN, F&I GM, 12-20 SF	0.21%	0.08%	0.41%	0.53%	0.59%	0.67%
46	MULTI-POST SIGN, F&I GM, 21-30 SF	0.03%	×	×	×	0.01%	0.01%
47	SINGLE POST SIGN, REMOVE	×	×	×	×	0.02%	0.02%
48	MULTI-POST SIGN, F&I GM, 31-50 SF	0.07%	0.32%	0.14%	0.16%	0.20%	0.23%
49	MULTI-POST SIGN, F&I GM, 51-100 SF	0.30%	0.51%	0.57%	0.23%	×	×
50	MULTI-POST SIGN, F&I GM, 101-200 SF	×	×	×	1.14%	×	×
51	MULTI-POST SIGN, REMOVE	×	×	×	×	0.03%	0.03%
52	OH STATIC SIGN STR, F&I, C 21-30 FT			×	×	×	×
53	OH STATIC SIGN STR, F&I, S 31-40 FT	×		×	×	×	×
54	RETRO-REFLECTIVE/RAISED PAVEMENT MARKERS	0.02%	0.01%	0.04%	0.03%	0.07%	0.06%
55	PAINTED PAVT MARK,STD,WHITE,SOLID,6"	×	×				
56	PAINTED PAVT MARK,STD,WHITE,SKIP, 6"	×	×				
57	THERMOPLASTIC, STD-OP, WHITE, SOLID, 6"						
58	THERMOPLASTIC, STD-OP, WHITE, SKIP, 6"						
59	INITIAL CONTINGENCY AMOUNT (DO NOT BID)	0.99%	0.99%	0.99%	0.99%	×	×
60	RURAL LIGHT POLES - 52 AT $10,000.00			×	×	×	×

Fig. 1. Cost items utilized in the study.

J. Constr. Eng. Manage., 2021, 147(3): 04020180

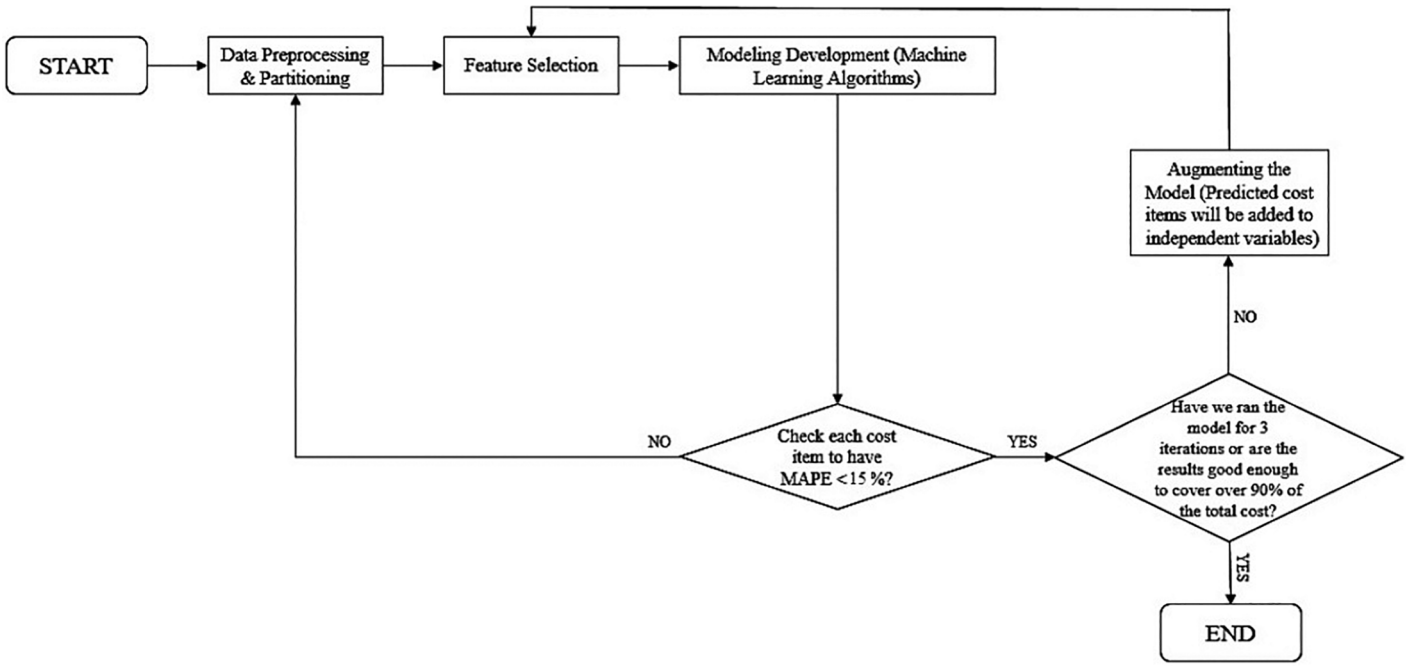

Fig. 2. Pipeline of this study.

monthly level throughout the entire analysis period. The highlighted cost items are the ones left out due to the lack of the monthly historical information. The ***x*** value in Fig. 1 indicates that the specific construction type does not have the associated cost item as one of the items calculating the total cost of the per-mile construction.

This study utilized five categories of independent variables (69 variables). The details of the predictors employed in this research are shown in Appendix I. Regarding the predictors, this study employed 14,076 data points for studying 69 variables over 204 months. The construction market category has 28 variables such as building permits and construction spending. Construction spending and housing permits represent the number of the construction market. The national highway construction cost index (NHCCI) was another critical candidate variable that has been employed for construction bids over the last years. Moreover, this study also employed Florida DOT disbursements and historical revenue information.

The socioeconomics category has nine variables. This project collected the income, household size, labor force, and length of the paved road to study the socioeconomic impact on the highway construction cost. The US economy category has 28 variables. Gross domestic product depicts the national income of the United States, and the consumer price index is broadly utilized to describe inflation at the federal level. Regarding stock market indexes, Dow Jones, NASDAQ index, and S&P 500 index were employed as they are widely evaluated as one of the preeminent potential factors for the US. The energy market predictors including crude oil price, natural gas price, gas price, and electricity price were used as a measure of energy price levels.

Model Development

The primary goal of this research was to forecast the total cost of each type of roadway construction expansion (either four- or six-lane construction or adding lanes to urban or rural roads). Fig. 2 represents the pipeline of this project. The pipeline of this work consists of data preprocessing, feature selection, model creation, parameter optimization, and evaluation using the Scikit-learn (Pedregosa et al. 2011) library for machine learning in the Python (Python 2020) programming language. During preprocessing, data was standardized and divided into training, testing, and validation datasets. Then, the validation and training datasets were fed to a feature selection module that selects the essential features within the data and removes other independent variables. This study examined multiple approaches for both feature selection and linear and nonlinear modeling. Through this process, the most appropriate model (for each cost item) was found.

After performing the modeling development in Run #1, the failed cost items that had an error (MAPE was be used as a measure) over 15% were isolated. The variables that were predicted correctly were added as independent variables to the pool of predictors, and the application was executed again. This approach demonstrates that an increase in the number of cost items can be accurately forecasted through this system.

Data Preprocessing and Partitioning

At the data preprocessing stage, all independent variables were converted to a number. Then, the numeric data were standardized to normal distributions with average of zero and standard deviation of one to support the regularizations of the models. Following standardization, the processed data were split into training, test, and validation.

As the data under study is a time series, the integrity and temporal continuity of the data were crucial components to examine; randomly splitting the dataset into several parts for validation would not be proper. Fig. 3 illustrates the nested cross-validation expanding window assessment approach that was used in this study. This study divided the historical cost information of the 17,121 projects based on their year of execution for the training, test, and validation datasets of each split. In this approach, the training set starts with the first 3 years, followed by the next 3 years as a validation set. To implement the expanding window

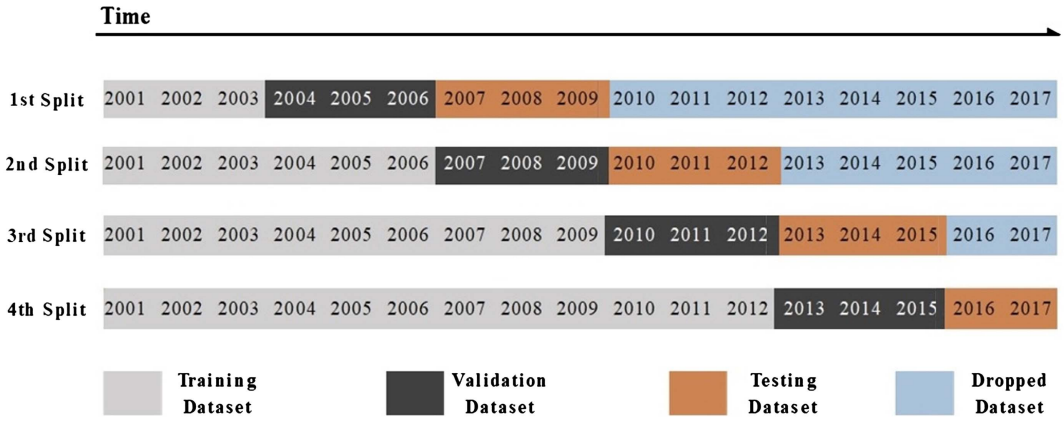

Fig. 3. Nested cross-validation; expanding window.

and cover the dataset, the training set was extended by 3 years in each phase. The testing dataset consisted of the next three successive years of the dataset after the validation dataset. Each split went through the pipeline of the study shown in Fig. 3. Using the results of each split, the error was then averaged. As depicted in Fig. 3, in each fold of cross-validation, the training, validation, and testing data points were different from each other. For example, in the fourth split, the projects' historical data from 2001 to 2012 was used to train the models, 2013 to 2015 to test, and 2016 and 2017 to validate them. As a result, the model is tested on several real project data costs from 2016 and 2017 in the presented case study. The final reported accuracy and models' details are all based on real project information and the models have never been trained and tested on the same datapoints. This method ensures that the final model is robust and is not an overfit model or a randomly accurate one.

All the models were trained on the training data and evaluated on the validation set. In order to find the optimum feature selection tool, model, and parameters, a grid search was performed in the parameter optimization stage using the validation set.

Performance Measurement Scales

To examine the performance of the feature selections and modeling approaches, four measures of error, including root mean squared error (RMSE), MAPE, coefficient of determination (R^2), and mean absolute error (MAE) were considered. In cases such as this study investigating the highway construction cost dataset, the aim was to produce the best possible forecast while understanding the possible error in those estimates. The RMSE and the MAPE were better measures of accuracy in this regard, as they provided insight into the possible error of the forecasts. Moreover, by using RMSE and MAPE, both scale-dependent and scale-independent measures were considered such that someone who has no idea what forms a significant error in the case study would also be able to perceive the results. However, similar to R^2, the RMSE is sensitive to occasional significant errors. Furthermore, the MAPE provides the most suitable mean of evaluating the error in this research, so the models were trained on the data and evaluated using this metric on the test set.

Feature Selection

Feature selection is the method of selecting the most appropriate predictors and eliminating unnecessary variables from the pool of potential predictors. Depending on the model's structure, feature selection can improve a model's accuracy. This method was carried out by finding the contribution of each variable to the models' precision and then eliminating unnecessary and repetitive variables while also maintaining the most beneficial ones. Standardization is necessary before feature selection because the independent variables have different magnitudes of order and using them as-is might result in the ones with small magnitudes being overlooked. For each parameter set, the cross-validation method discussed earlier served to train, validate, and test the model. In this work, three methods were employed to decide the main features. First, valuable features were chosen via a model [SelectFromModel function from Scikit-learn (Pedregosa et al. 2011)]. Various modeling techniques capable of implicit feature selection, including random forest (RF) regression, ridge regression (Ridge), Bayesian ridge (BR) regression, and decision tree (DT), were employed in this section. The importance threshold considered for the selection parameter of this step changes between 0.25, 0.5, 0.75, 1, 1.25, 1.5, and 1.75. Second, the recursive feature elimination (RFE) [as used in Scikit Learn (Pedregosa et al. 2011)] was conducted. In this process, the least essential features are eliminated recursively until the most appropriate features are found. The models used to determine the importance of features are the same as the previous step (RFE-RF, RFE-Ridge, RFE-BR, and RFE-DT). In the RFE step, the number of ultimately selected features varies between 1, 3, 5, 10, 20, 30, 40, 50, and 60. Third, a scoring function was used to find the K best features in the dataset [SelectKBest in Scikit Learn (Pedregosa et al. 2011)]. The scoring functions used in this work were ANOVA F-value (FCLASSIF) and mutual information (MFCLASSIF). The number of ultimately selected features of this step also fluctuates between 1, 3, 5, 10, 20, 30, 40, 50, and 60. These feature selection approaches were implemented inside a grid search and finally compared to find the best set of parameters.

Modeling Approaches

Multiple machine learning algorithms were employed in this study, particularly those based on nonlinear relationships between variables to forecast the cost items. The models [SelectFromModel function from Scikit-learn (Pedregosa et al. 2011)] that were used in this study included decision tree, random forest, K-nearest neighbors (KNN), and neural network. Moreover, linear regression models, including linear regression (Linear), ridge regression, Bayesian ridge, stochastic gradient descent (SGD) regression, and passive-aggressive (PA) regression were evaluated to find the level of improvement of using nonlinear models. This selection of models

Dependent Variables	Predicted in Stage 1	Predicted in Stage 2	Unable to Predict
MAINTENANCE OF TRAFFIC	✓		
MOBILIZATION	✓		
SEDIMENT BARRIER	✓		
FLOATING TURBIDITY BARRIER	✓		
STAKED TURBIDITY BARRIER- NYL REINF PVC		✓	
SOIL TRACKING PREVENTION DEVICE	✓		
LITTER REMOVAL	✓		
MOWING	✓		
CLEARING & GRUBBING	✓		
REGULAR EXCAVATION		✓	
BORROW EXCAVATION, TRUCK MEASURE			✓
EMBANKMENT		✓	
TYPE B STABILIZATION	✓		
OPTIONAL BASE,BASE GROUP 04		✓	
OPTIONAL BASE,BASE GROUP 09		✓	
MILLING EXIST ASPH PAVT, 1" AVG DEPTH	✓		
MILLING EXIST ASPH PAVT, 4" AVG DEPTH	✓		
SUPERPAVE ASPH CONC, TRAF C,PG76-22,PMA	✓		
SUPERPAVE ASPH CONC, TRAF D,PG76-22,PMA		✓	
ASPH CONC FC,INC BIT,FC-5,PG76-22,PMA		✓	
ASPH CONC FC,TRAFFIC B,FC-9.5,PG 76-22		✓	
CONC CLASS II, ENDWALLS		✓	
INLETS, DT BOT, TYPE D		✓	
INLETS, DT BOT, TYPE E		✓	
INLETS, BARRIER WALL	✓		

Dependent Variables	Predicted in Stage 1	Predicted in Stage 2	Unable to Predict
MANHOLES, J-7			✓
PIPE CULV, OPT MATL, ROUND,24"SD	✓		
PIPE CULV, OPT MATL, ROUND, 24"S/CD	✓		
PIPE CULV, OPT MATL, ROUND, 30"S/CD	✓		
PIPE CULV, OPT MATL, ROUND, 36"S/CD	✓		
PIPE CULV, OPT MATL, ROUND, 42"S/CD		✓	
PIPE CULV, OPT MATL, ROUND, 54"S/CD		✓	
MITERED END SECT, OPTIONAL RD, 24" SD	✓		
MEDIAN CONC BARRIER WALL	✓		
SHLDR CONC BARRIER, RIGID-SHLDR	✓		
CONCRETE DITCH PAVT, NR, 3"	✓		
RUMBLE STRIPS, GROUND-IN, 16" MIN. WIDTH	✓		
FENCING, TYPE B, 5.1-6.0', STANDARD	✓		
PERFORMANCE TURF		✓	
PERFORMANCE TURF, SOD	✓		
SINGLE POST SIGN, F&I GM	✓		
SINGLE POST SIGN, F&I GM, 12-20 SF	✓		
MULTI- POST SIGN, F&I GM, 21-30 SF	✓		
SINGLE POST SIGN, REMOVE	✓		
MULTI- POST SIGN, F&I GM, 31-50 SF	✓		
MULTI- POST SIGN, F&I GM, 51-100 SF	✓		
MULTI- POST SIGN, F&I GM, 101-200 SF	✓		
MULTI- POST SIGN, REMOVE		✓	
RETRO-REFLECTIVE/RAISED PAVEMENT MARKERS	✓		
INITIAL CONTINGENCY AMOUNT (DO NOT BID)			✓

Fig. 4. Prediction stage of each cost item.

allows the user to compare models with different levels of linearity or nonlinearity, as well as having control over parametric models.

For the RF and DT algorithms in this research, the model parameter (MP), which is the maximum depth of the trees, varies between 5, 20, 50, 75, 100, and 200. Regarding the *K*-nearest neighbors algorithm employed, the model parameter (number of neighbors, *K*) changes between 1, 3, 5, 7, 10, and 16. Concerning the neural network models, the MP, which represents the hidden layer size (number of neurons), varies between 1, 2, and 4. Conversely, in the linear algorithms, for the ridge regression, the MP represents the

Stage 1 Results					Validation Dataset		Test Dataset	
Label Name	Feature Selection Approach	Model	SP	MP	R^2	MAPE	R^2	MAPE
MAINTENANCE OF TRAFFIC	RF	Ridge	1	0.1	0.99	1.78%	0.72	8.39%
MOBILIZATION	RF	Linear	1	N/A	0.89	2.60%	0.73	10.01%
SEDIMENT BARRIER	RFERF	PA	20	1	1.00	1.14%	0.84	9.21%
FLOATING TURBIDITY BARRIER	MFCLASSIF	PA	30	1	0.97	3.31%	0.69	6.78%
STAKED TURBIDITY BARRIER- NYL REINF PVC	RFEBaysianRidge	PA	20	1	0.99	5.35%	0.20	15.91%
SOIL TRACKING PREVENTION DEVICE	RFERidge	BaysianRidge	10	0.1	0.99	1.38%	0.69	9.00%
LITTER REMOVAL	RFERF	BaysianRidge	30	1	1.00	1.23%	0.70	8.27%
MOWING	MFCLASSIF	BaysianRidge	20	0.1	0.99	1.25%	0.60	9.18%
CLEARING & GRUBBING	RFEBaysianRidge	BaysianRidge	10	0.1	1.00	1.48%	0.66	9.86%
REGULAR EXCAVATION	RFEBaysianRidge	Ridge	3	1	0.97	3.74%	0.39	16.39%
BORROW EXCAVATION, TRUCK MEASURE	BaysianRidge	Ridge	10	10	0.55	8.20%	0.32	25.30%
EMBANKMENT	RFERidge	PA	5	100	0.99	3.70%	0.55	17.21%
TYPE B STABILIZATION	RFERF	BaysianRidge	50	0.1	1.00	0.28%	0.91	3.00%
OPTIONAL BASE,BASE GROUP 04	BaysianRidge	BaysianRidge	1.75	0.1	0.97	3.20%	0.54	15.03%
OPTIONAL BASE,BASE GROUP 09	MFCLASSIF	BaysianRidge	30	0.1	0.98	1.18%	0.48	16.60%
MILLING EXIST ASPH PAVT, 1" AVG DEPTH	RFERidge	BaysianRidge	5	10000	0.99	0.93%	0.76	6.47%
MILLING EXIST ASPH PAVT, 4" AVG DEPTH	MFCLASSIF	BaysianRidge	40	0.1	1.00	0.78%	0.86	6.60%
SUPERPAVE ASPH CONC, TRAF C, PG76-22,PMA	MFCLASSIF	BaysianRidge	20	1	0.99	0.33%	0.74	3.77%
SUPERPAVE ASPH CONC, TRAF D, PG76-22,PMA	MFCLASSIF	BaysianRidge	20	10	0.99	4.71%	0.43	15.49%
ASPH CONC FC,INC BIT,FC-5,PG76-22,PMA	RF	Linear	1	N/A	0.99	2.79%	0.52	18.99%
ASPH CONC FC,TRAFFIC B,FC-9.5,PG 76-22	RFERidge	NN	1	4	0.98	2.59%	0.49	15.14%
CONC CLASS II, ENDWALLS	MFCLASSIF	BaysianRidge	5	1	0.98	3.30%	0.61	15.89%
INLETS, DT BOT, TYPE D, <10'	RF	BaysianRidge	0.25	100	0.99	3.50%	0.39	15.21%
INLETS, DT BOT, TYPE E, <10'	RFERidge	NN	5	4	0.99	5.83%	0.55	23.20%
INLETS, BARRIER WALL, <=10'	Ridge	BaysianRidge	1.5	0.1	0.97	1.83%	0.71	8.84%

Stage 1 Results					Validation Dataset		Test Dataset	
Label Name	Feature Selection Approach	Model	SP	MP	R^2	MAPE	R^2	MAPE
MANHOLES, J-7, <10'	Ridge	BaysianRidge	1.75	0.1	0.99	4.32%	0.60	17.75%
PIPE CULV, OPT MATL, ROUND,24"SD	BaysianRidge	PA	1	10	0.99	0.60%	0.80	4.21%
PIPE CULV, OPT MATL, ROUND, 24"S/CD	Ridge	PA	0.75	0.1	0.99	2.07%	0.65	11.02%
PIPE CULV, OPT MATL, ROUND, 30"S/CD	RFEDT	BaysianRidge	20	10	0.99	1.41%	0.51	10.81%
PIPE CULV, OPT MATL, ROUND, 36"S/CD	RFERidge	PA	30	10	0.99	1.50%	0.85	6.06%
PIPE CULV, OPT MATL, ROUND, 42"S/CD	RFEDT	Ridge	10	1	0.99	4.29%	0.09	15.56%
PIPE CULV, OPT MATL, ROUND, 54"S/CD	MFCLASSIF	PA	30	10000	0.99	3.84%	0.26	16.20%
MITERED END SECT, OPTIONAL RD, 24" SD	RFERF	BaysianRidge	40	1E+06	1.00	0.40%	0.85	5.85%
MEDIAN CONC BARRIER WALL	MFCLASSIF	BaysianRidge	3	10	0.99	0.70%	0.31	5.47%
SHLDR CONC BARRIER, RIGID-SHLDR	Ridge	BaysianRidge	1.75	1	0.97	2.31%	0.81	14.34%
CONCRETE DITCH PAVT, NR, 3"	Ridge	BaysianRidge	1.75	1	0.99	0.68%	0.55	6.17%
RUMBLE STRIPS, GROUND-IN, 16" MIN. WIDTH	RFEBaysianRidge	BaysianRidge	5	0.1	0.99	0.76%	0.80	4.91%
FENCING, TYPE B, 5.1-6.0', STANDARD	RFERF	BaysianRidge	10	1E+06	0.99	0.48%	0.53	4.00%
PERFORMANCE TURF	RFERF	Linear	10	N/A	0.99	7.11%	0.58	17.99%
PERFORMANCE TURF, SOD	RFEBaysianRidge	BaysianRidge	10	1	0.99	0.59%	0.86	3.85%
SINGLE POST SIGN, F&I GM, <12 SF	DT	BaysianRidge	0.25	1	1.00	0.22%	0.93	2.74%
SINGLE POST SIGN, F&I GM, 12-20 SF	MFCLASSIF	BaysianRidge	3	0.1	1.00	0.55%	0.86	4.89%
MULTI- POST SIGN, F&I GM, 21-30 SF	BaysianRidge	PA	1.75	1E+06	1.00	0.39%	0.90	3.42%
SINGLE POST SIGN, REMOVE	BaysianRidge	Linear	0.75	N/A	0.99	0.68%	0.93	4.44%
MULTI- POST SIGN, F&I GM, 31-50 SF	BaysianRidge	Ridge	1.75	0.1	1.00	0.48%	0.93	4.44%
MULTI- POST SIGN, F&I GM, 51-100 SF	RFEBaysianRidge	Ridge	10	0.1	1.00	0.46%	0.77	6.53%
MULTI- POST SIGN, F&I GM, 101-200 SF	BaysianRidge	Linear	0.5	N/A	0.98	1.12%	0.96	2.85%
MULTI- POST SIGN, REMOVE	RF	Linear	0.25	N/A	1.00	4.45%	0.24	17.16%
RETRO-REFLECTIVE/RAISED PAVEMENT MARKERS	BaysianRidge	BaysianRidge	1.75	1	0.99	0.58%	0.66	2.58%
INITIAL CONTINGENCY AMOUNT, DO NOT BID	RFEBaysianRidge	NN	50	4	0.93	5.61%	0.34	15.40%

Fig. 5. Results of Stage 1 of the modeling process.

regularization strength (alpha) and varies between 0.1, 1, 10, 100, 10,000, and 1×10^6; for Bayesian ridge regression, the model parameter shows the shape and inverse scale parameters of the prior gamma distribution (alpha_1 and alpha_2) and varies between 0.1, 1, 10, 100, 10,000, and 1×10^6. Regarding the stochastic gradient descent regression, the MP represents the elastic net mixing parameter of L1 and L2 regularization (L1 ratio) and fluctuates between 0, 0.15, 0.3, 0.5, 0.75, and 1. Ultimately, regarding passive-aggressive regression, MP shows the maximum step size (Regularization C) and changes between 0.1, 1, 10, 100, 10,000, and 1×10^6.

As demonstrated in the feature selection and the modeling approach sections, the developed hyperparameter optimization includes a wide range of values for the parameters from reasonably low values to reasonably high values, so that it could be applied to various datasets with differing characteristics.

Results and Discussion

To test the feasibility of this approach, the FDOT highway construction cost data between 2001 and 2017 were utilized. These cost items covered about 92.6% of the average total cost of highway construction. Among the 50 cost items (dependent variables) in this study that we collected from the historical data from FDOT, 32 cost items were predicted in Stage 1 of the analysis, and 15 cost items were predicted in the second stage, and three were not predicted with high accuracy (MAPE below 15%). Fig. 4 shows the stage each cost item was predicted. Moreover, the three cost items that were not predicted accurately are depicted.

Fig. 5 represents the results of the first stage of running the inputs through the study pipeline on both validation and test dataset. Moreover, the optimized feature selection approach and modeling approach with their selection parameter (SP) and modeling parameter (MP) are depicted in Fig. 5. In general, the linear models performed better than the nonlinear models. At this stage, we could successfully forecast 32 cost items with more than 85% accuracy. The highlighted cost items in Fig. 5 are the ones with higher than 15% forecast error and were chosen to move to the second stage. In the second stage, the cost items that were successfully predicted in the first stage were employed as supplemental inputs (predictors) for the second stage.

Fig. 6 illustrates the results of the second stage of the modeling process. With the increased pool of dependent variables, we could forecast 15 out of the 18 cost items that initially had higher than the

Stage 2 Results					Validation Dataset		Test Dataset	
Label Name	Feature Selection Approach	Model	SP	MP	R^2	MAPE	R^2	MAPE
STAKED TURBIDITY BARRIER- NYL REINF PVC	MFCLASSIF	Linear	50	N/A	1	2.04%	0.56	13.69%
REGULAR EXCAVATION	RF	NN	0.5	4	1	0.89%	0.59	7.51%
BORROW EXCAVATION, TRUCK MEASURE	RF	NN	60	N/A	0.65	5.20%	0.44	18.30%
EMBANKMENT	RFE-Bayesian-Ridge	PA	30	10	1	0.96%	0.75	7.68%
OPTIONAL BASE, BASE GROUP 04	Ridge	Linear	0.75	N/A	1	0.22%	0.88	2.04%
OPTIONAL BASE, BASE GROUP 09	DT	Ridge	1.25	10	0.99	0.68%	0.57	5.92%
SUPERPAVE ASPH CONC, TRAF D, PG76-22, PMA	RFE-Bayesian-Ridge	Bayesian Ridge	40	1	1	0.48%	0.78	8.29%
ASPH CONC FC, INC BIT, FC-5, PG76-22, PMA	RFE-Ridge	SGD	5	1	0.99	0.54%	0.83	2.58%
ASPH CONC FC, TRAFFIC B, FC-9.5, PG 76-22	RFERF	Ridge	5	100	0.99	0.39%	0.3	1.60%
CONC CLASS II, ENDWALLS	MFCLASSIF	PA	30	1	1	0.25%	0.86	2.20%
INLETS, DT BOT, TYPE D, <10'	Ridge	PA	1.75	1E+06	1	0.55%	0.83	3.40%
INLETS, DT BOT, TYPE E, <10'	MFCLASSIF	NN	1	2	1	1.99%	0.76	14.06%
MANHOLES, J-7, <10'	Bayesian-Ridge	Linear	0.5	N/A	0.99	2.88%	0.55	16.81%
PIPE CULV, OPT MATL, ROUND, 42"S/CD	Ridge	Linear	0.25	N/A	1	0.63%	0.86	5.16%
PIPE CULV, OPT MATL, ROUND, 54"S/CD	Bayesian-Ridge	PA	1.25	10	1	0.86%	0.58	11.95%
PERFORMANCE TURF	Bayesian-Ridge	Linear	1	N/A	1	1.65%	0.82	14.21%
MULTI- POST SIGN, REMOVE	Bayesian-Ridge	Linear	0.75	N/A	1	1.32%	0.66	10.20%
INITIAL CONTINGENCY AMOUNT, DO NOT BID	RFE-Ridge	PA	20	1	0.94	4.42%	0.47	16.36%

Fig. 6. Results of Stage 2 of the modeling process.

Table 1. Algorithm approaches of the developed model

Models	No. predicted cost items	Constructing new urban 6L (%)	Constructing new urban 4L (%)	Constructing new rural 6L (%)	Constructing new rural 4L (%)	Widening 6L to 8L (%)	Widening 4L to 6L (%)	Average (%)
Linear models	45	92.16	90.97	94.01	93.25	84.72	82.93	89.67
Linear	8	18.67	20.28	14.10	13.27	12.74	13.23	15.38
Ridge	5	14.50	13.34	16.46	13.65	13.96	14.68	14.43
Bayesian ridge	21	48.42	41.85	43.78	42.80	44.30	41.50	43.78
Stochastic gradient descent	1	1.92	1.41	0.00	0.08	6.92	5.86	2.70
Passive-aggressive	10	8.65	14.09	19.67	23.45	6.80	7.66	13.39
Nonlinear models	2	1.81	1.96	4.26	5.00	2.12	2.39	2.92
Neural network	2	1.81	1.96	4.26	5.00	2.12	2.39	2.92

Note: 4L = four lane; 6L = six lane; and 8L = eight lane.

Table 2. Categorical feature importance of the developed model

Categorical feature importance	Construction market (%)	Socioeconomic (%)	Temporal (%)	US economy (%)	Energy market (%)
Constructing new urban 6L	81.26	7.88	4.95	4.20	1.70
Constructing new urban 4L	79.95	8.51	5.50	4.38	1.66
Constructing new rural 6L	79.05	5.34	6.84	5.98	2.79
Constructing new rural 4L	78.68	5.47	6.97	5.76	3.11
Widening 6L to 8L	81.74	5.43	5.30	5.26	2.26
Widening 4L to 6L	81.23	5.41	5.53	5.53	2.29
Average all construction types	80.32	6.34	5.85	5.19	2.30

Note: 4L = four lane; 6L = six lane; and 8L = eight lane.

Table 3. Important cost items of the developed model on validation datasets for the fourth split

Cost item	Average percentage impact in all six methods (%)	Feature selection approach	Modeling approach	Selection parameter (SP)	Model parameter (MP)	MAPE (validation dataset, fourth split) (%)	MAPE (test dataset, fourth split) (%)
SHLDR CONC BARRIER, RIGID-SHLDR	17.54	Ridge	Bayesian ridge	1.75	1	0.59	12.42
SUPERPAVE ASPH CONC, TRAF C, PG76-22	10.39	MFCLASSIF	Bayesian ridge	20	1	0.25	2.98
MAINTENANCE OF TRAFFIC	6.88	RF	Ridge	1	0.1	1.36	7.97

threshold error (15% MAPE on the test dataset). Overall, this processing system resulted in forecasting 47 out of 50 cost items under the study with more than 92% on average accuracy on various highway construction types.

The summary of the result of the developed model is presented in Table 1. It is evident that linear models outperform the nonlinear algorithms within the scope of this study. Among 47 predicted cost items in this study, 45 cost items (about 89.6% coverage of the total highway construction cost) were predicted by linear models. Only two cost items were predicted by nonlinear algorithms, covering about 2.92% of total cost of the highway construction. Within the various linear models examined in this study, Bayesian ridge performed better for 21 cost items (out of 45 items predicted by linear models) covering about 43.78% of the total cost. Moreover, concerning nonlinear models, the NN algorithm was able to predict "REGULAR EXCAVATION," and "INLETS, DT BOT, TYPE E, <10′" with higher accuracy compared to the linear models.

The categorical feature importance of various construction types is depicted in Table 2. On average, the construction market category, with 80.32%, had the most significant impact on the highway construction cost prediction model, while the socioeconomic category with 6.4% was second. Additionally, the US economy had 5.19%, energy market had 2.3%, and temporal predictors had 5.85% importance. The categorical feature importance of all 47 predicted cost items in this study is shown in Appendix II.

Three critical cost items that had high average percentage impact on the various construction types studied as shown in Table 3 and were selected as a sample for further in-depth analysis. In the previous sections, the reported MAPE on the validation and test datasets were obtained from the average MAPE over all the folds for each cost item. To observe the effect of model parameters on its performance, the results on the fourth split of the data are analyzed. This split (fourth split, consisting of 12 years of training, 3 years of validation, and 2 years of testing dataset) of the nested cross-validation covers all the datasets and outperformed the other three folds for each cost item.

Fig. 7 depicts the model optimization of the shoulder concrete barrier, "SHLDR CON BARRIER" on the fourth split of the dataset. The feature selection approach of Ridge with a selection parameter of 1.75 and the Bayesian ridge model algorithm with a model parameter of 1 has the lowest MAPE of 0.59% on the validation dataset.

Fig. 8 shows the categorical and individual feature importance of the "SHLDR CONC BARRIER." The construction market

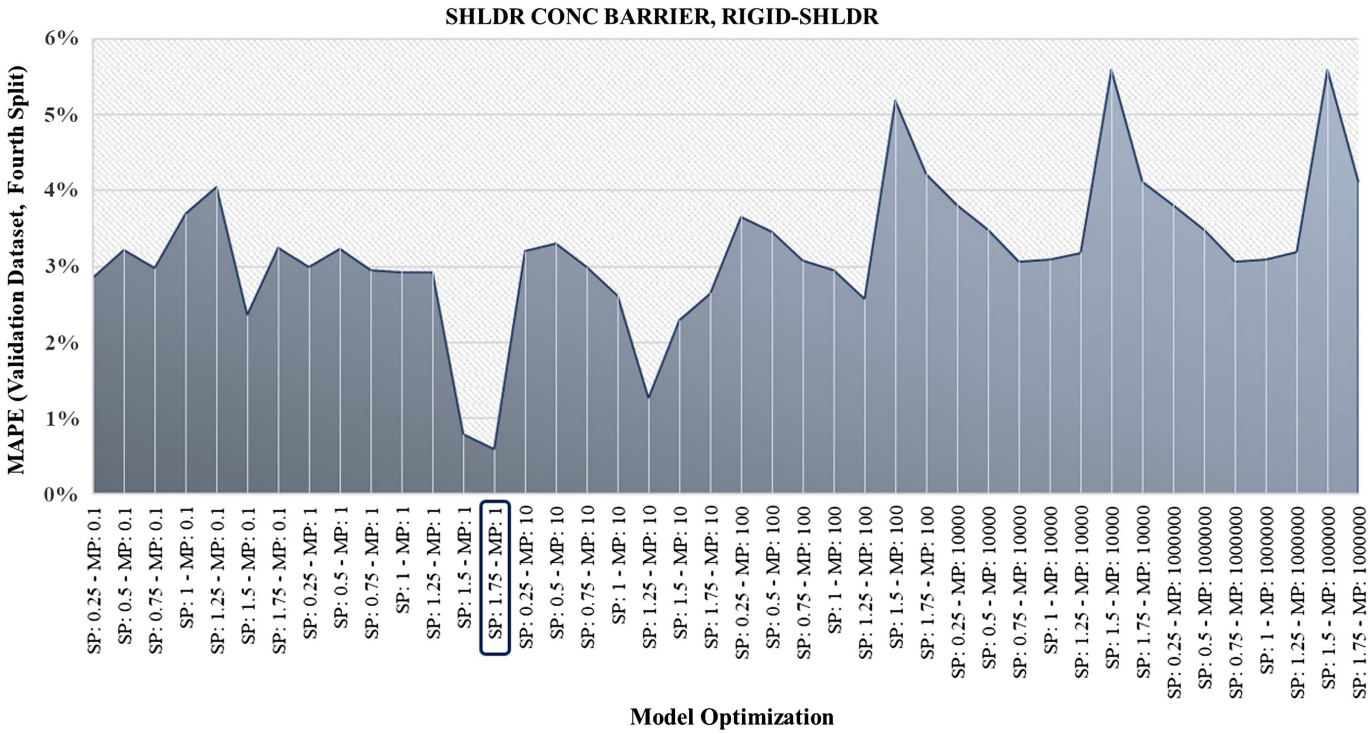

Fig. 7. Model optimization of validation dataset for "SHLDR CONC BARRIER" on the fourth split.

category had the highest impact with 89.6% importance on this highway construction cost item's prediction model. Socioeconomic variables with 10% importance had the second rank for this cost item. Moreover, the NHCCI predictor played a significant role in predicting this cost item.

The following equation shows all the forecasting formula for the "SHLDR CONC BARRIER" cost item:

$$\begin{aligned}\text{Cost of "SHLDR CONC BARRIER, RIGID-SHLDR"} &= 0.824025525 \times \text{"NHCCI Global"} + 0.629230173 \times \text{"Other Roads PRODUCT AREAS"} + 0.570224754 \times \text{"Right of Way"} \\ &+ 0.568821925 \times \text{"Local Government Grants"} + 0.374852392 \times \text{"State Motor Vehicle Tax"} + 0.240822376 \times \text{"Bond Retirement"} \\ &+ 0.200928502 \times \text{"HHEUS"} + 0.191619513 \times \text{"Other State Funding"} + 0.178818362 \times \text{"AHEPNECUS"} + 0.100279817 \\ &\times \text{"State Motor Fuel Tax"} + 0.046363366 \times \text{"NPHUABPFL"} + 0.018570434 \times \text{"DJI"} + 0.014885667 \times \text{"URUS"}\end{aligned}$$

Fig. 9 depicts the model optimization of the "ASPH CONC, TRAF C" on the fourth split of the dataset for this cost item. The feature selection approach of *MFCLASSIF* with a selection parameter of 20 and the Bayesian ridge model algorithm with a model parameter of 1 has the lowest MAPE of 1.36% on the validation dataset.

Fig. 10 depicts the categorical and individual feature importance of the Superpave asphalt concrete "Superpave ASPH CONC, TRAF C." The construction market had the highest impact, with 89.6% importance on this highway construction cost item's prediction model. Temporal variables with 10.9% importance had the second rank for this cost item. Socioeconomic variables had the third rank of importance with 5% importance level. Lastly, the NHCCI predictor played a key role in predicting this cost item.

The following equation shows all the important predictors and their coefficients on the standardized dataset for the "SUPERPAVE ASPH CONC, TRAF C":

$$\begin{aligned}\text{Cost "SUPERPAVE ASPH CONC, TRAF C, PG76-22, PMA"} &= 0.748396041 \times \text{"NHCCI Global"} + 0.267033067 \times \text{"Local Government Grants"} + 0.23817534 \times \text{"Bond Retirement"} \\ &+ 0.142771839 \times \text{"Number of Months from Beginning"} + 0.105815124 \times \text{"Administration"} + 0.092030559 \times \text{"YEAR"} \\ &+ 0.066819567 \times \text{"Interest"} + 0.064595363 \times \text{"State Motor Vehicle Tax"} + 0.062800303 \times \text{"Length Paved Roads Lane Miles"} \\ &+ 0.06138327 \times \text{"Total Florida State Revenue Sources"} + 0.05549031 \times \text{"Legislative Budget Request Amounts"} + 0.05273566 \\ &\times \text{"Capital Expenditures"} + 0.051153537 \times \text{"Total State Highway System(SHS)PRODUCT AREAS"} + 0.044435848 \times \text{"CLFFL"} \\ &+ 0.031637778 \times \text{"FederalFunding"} + 0.023584501 \times \text{"Tolls"} + 0.015849805 \times \text{"Total Florida State DOT Disbursements"} \\ &+ 0.01339556 \times \text{"Other State Funding"} + 0.010618173 \times \text{"GDP"} + 0.005356619 \times \text{"M2"}\end{aligned}$$

Fig. 11 depicts the model optimization of the "MAINTENANCE OF TRAFFIC" on the fourth split of the dataset for this cost item. The feature selection approach of RF with a selection parameter of 1 and the Ridge model algorithm with a model parameter of 0.1 had the lowest MAPE of 0.25% on the validation dataset.

Fig. 12 shows the categorical and individual feature importance of the "MAINTENANCE OF TRAFFIC." The construction market category of the variables had the highest impact with 89.59% importance on this highway construction cost item's prediction model. Temporal variables with 8.89% importance had the second rank for this cost item. Additionally, the right of way revenue stream of FDOT predictor played a key role in predicting this cost item.

The following equation shows the all the important predictors and their coefficients on the standardized dataset for the "MAINTENACE OF TRAFFIC":

$$\begin{aligned}\text{Cost "MAINTENANCE OF TRAFFIC"} &= 1.138962322 \times \text{"Right of Way"} + 0.847093766 \times \text{"Maintenance"} + 0.705406042 \times \text{"State Motor Fuel Tax"} \\ &+ 0.593492872\text{"NHCCI Global"} + 0.406601977 \times \text{"Number of Months from Beginning"} + 0.303016613 \times \text{"Interest"} \\ &+ 0.254637769 \times \text{"Total State Highway System(SHS) PRODUCT AREAS"} + 0.160810353 \times \text{"CEFL"} + 0.115055866 \\ &\times \text{"Total Florida State Revenue Sources"} + 0.044699378 \times \text{"BPLRUS"} + 0.028827425 \times \text{"CANUSER"} + 0.012286083 \\ &\times \text{"AECHCEUS"}\end{aligned}$$

Table 4 summarizes the coverage and accuracy of the results. All of the studied highway construction cost types had more than 90.95% accuracy with a minimum of 85.32% cost coverage and a maximum of 98.27%. As a result, we can argue that the results of the FDOT case study show the viability of this approach. The model was not capable of forecasting the cost item "BORROW EXCAVATION, TRUCK MEASURE" with an accuracy higher than 85%. This item has a 10.76% weight factor of the total cost of "widening 6 to 8 cost per-mile" and 12.10% of "widening of 4 to 6 cost per-mile," so the coverage of the cost per mile of the models for these two construction types was less than other types.

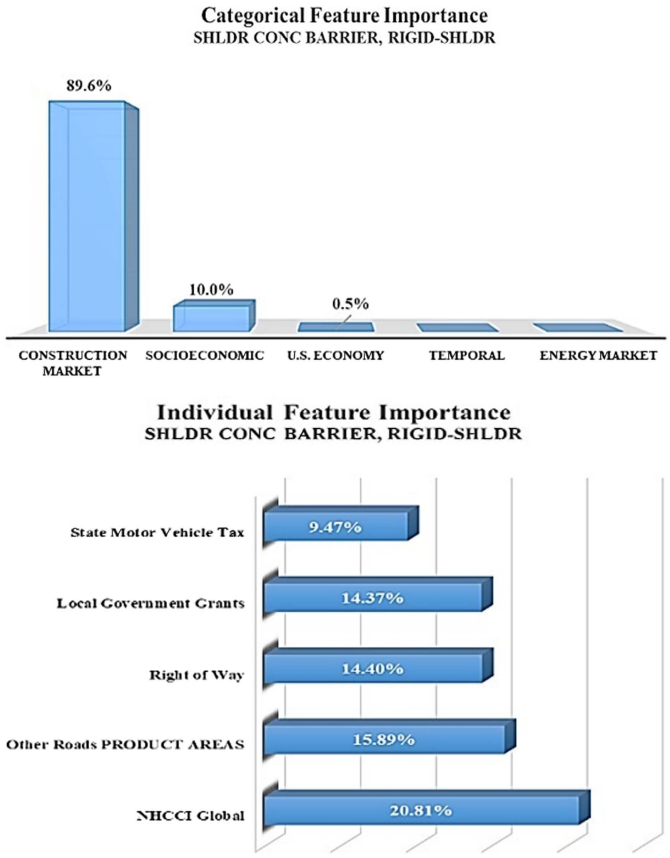

Fig. 8. Categorical and individual feature importance of "SHLDR CONC BARRIER."

Conclusion

An accurate estimate of the cost of infrastructure projects plays a vital role in their success. The existing literature lacks a cost prediction model that can automate the process and forecasts cost with a high degree of accuracy. The literature revealed that most studies focus on one or two algorithms, some linear and some nonlinear. Their results showed linear models versus nonlinear models worked best only on a case by case basis. This suggests the relationship between local and global variables for forecasting highway construction costs can be linear or nonlinear depending on the location and the specific issues such as type of projects or the level of analysis. As a result, a universal and generalizable framework that can optimize a cost forecasting model based on the characteristics of the input data is needed. In order to develop a universal and generalizable framework, this study employed a wide range of feature selection and modeling approaches (inclusive of all the methods reviewed from the literature) with a broad set of possible values and characteristics mentioned in the methodology. This ensures that every reader can easily insert their input data and find the best set of feature selection and modeling approaches characteristics to identify the leading factors in forecasting highway construction costs regardless of their location, specific project type, or study scope.

This study generated a framework for accurately predicting the total construction cost of highways at any given time in the future by employing machine learning. The pipeline of this work consisted of data preprocessing, feature selection, model creation (consisting of various linear and nonlinear algorithms), parameter optimization, and evaluation, all of which automate the cost prediction processes therefore reducing the number of manual activities and the need for skilled data scientists. One point of departure of this study from previous studies is the use of monthly highway construction cost data. Using the monthly data rather than annual data improved the model with more data points, resulting in cost prediction with higher accuracy. To test the model, our work used the FDOT highway construction cost items between 2001 and 2017. Additionally, we utilized five categories of independent variables (69 variables), including construction market variables,

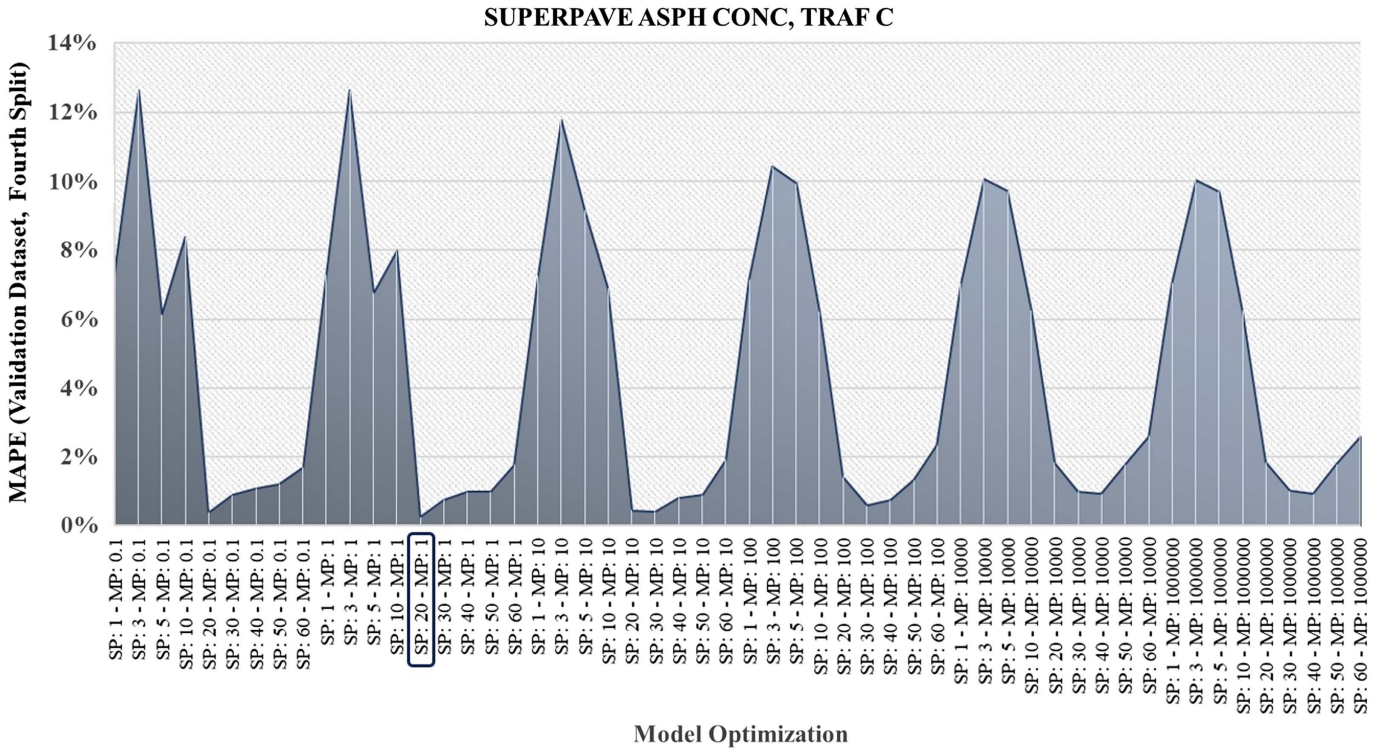

Fig. 9. Model optimization on the validation dataset for "SUPERPAVE ASPH CONC TRAF C" on fourth split.

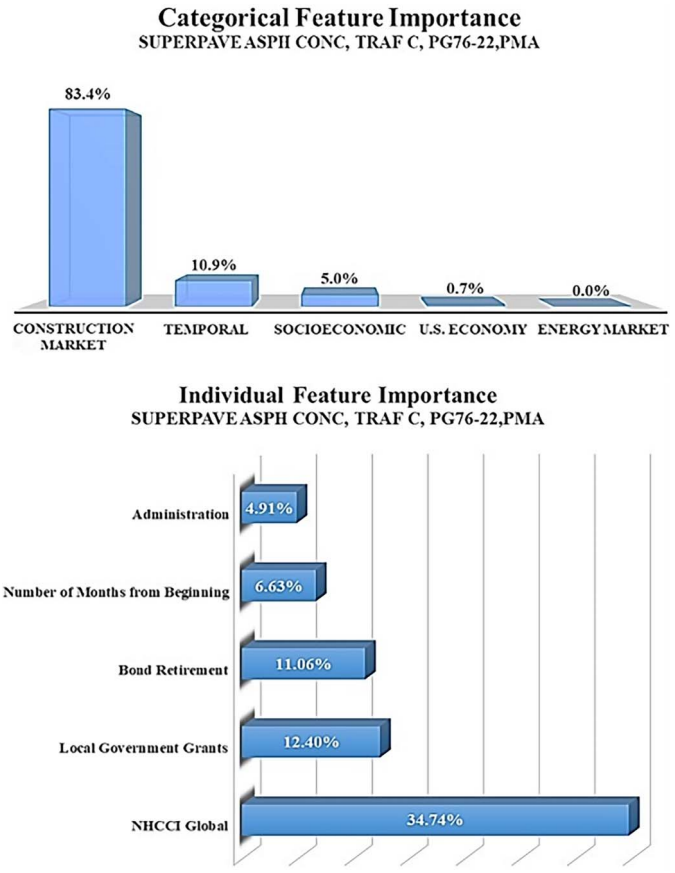

Fig. 10. Categorical and individual feature importance of "SUPERPAVE ASPH CONC TRAF C."

energy market variables, socioeconomics variables, and US economy variables.

From the 60 cost items (dependent variables) covering 100% of the total cost of six highway expansion types (constructing and widening), 10 cost items' monthly historical data were not available (about 7.4% of the total cost). A total of 32 cost items were predicted in Stage 1 of the analysis, 15 cost items were predicted in the second stage, and three were not predicted with high accuracy (MAPE below 15%). The model developed in this study covers 92.6% (on average) of the highway total cost per mile, 89.68% of which were predicted with linear models, while 2.92% utilized nonlinear algorithms. The highway prediction accuracy model developed in this study forecasted the FDOT highway cost with 92.51% accuracy (on average among different types). The results of the study show that the construction market category of the variables with 80.32% had the highest impact on the highway construction cost forecast, while the socioeconomic category with 6.4% was second. Additionally, the US economy had a 5.19% impact, energy market had 2.3%, and temporal predictors had 5.85%.

This study used a data-driven approach to find the best features and modeling approach. All possible linear and nonlinear models and the various possible values for parameters and independent variables in both the feature selection and modeling approach selection were employed to generate a comprehensive framework so that any user can follow the steps and employ this framework. In doing so, they can find the best feature selection method, the most appropriate set of features, and the best algorithms customized to their specific dataset. This claim is based on the fact that the developed framework not only includes all the approaches implemented in the reviewed literature but also goes beyond them in terms of the number and complexity of the algorithms and feature selection methods. Furthermore, this study employed 17 years of the historical cost data of 17,121 projects of various size to

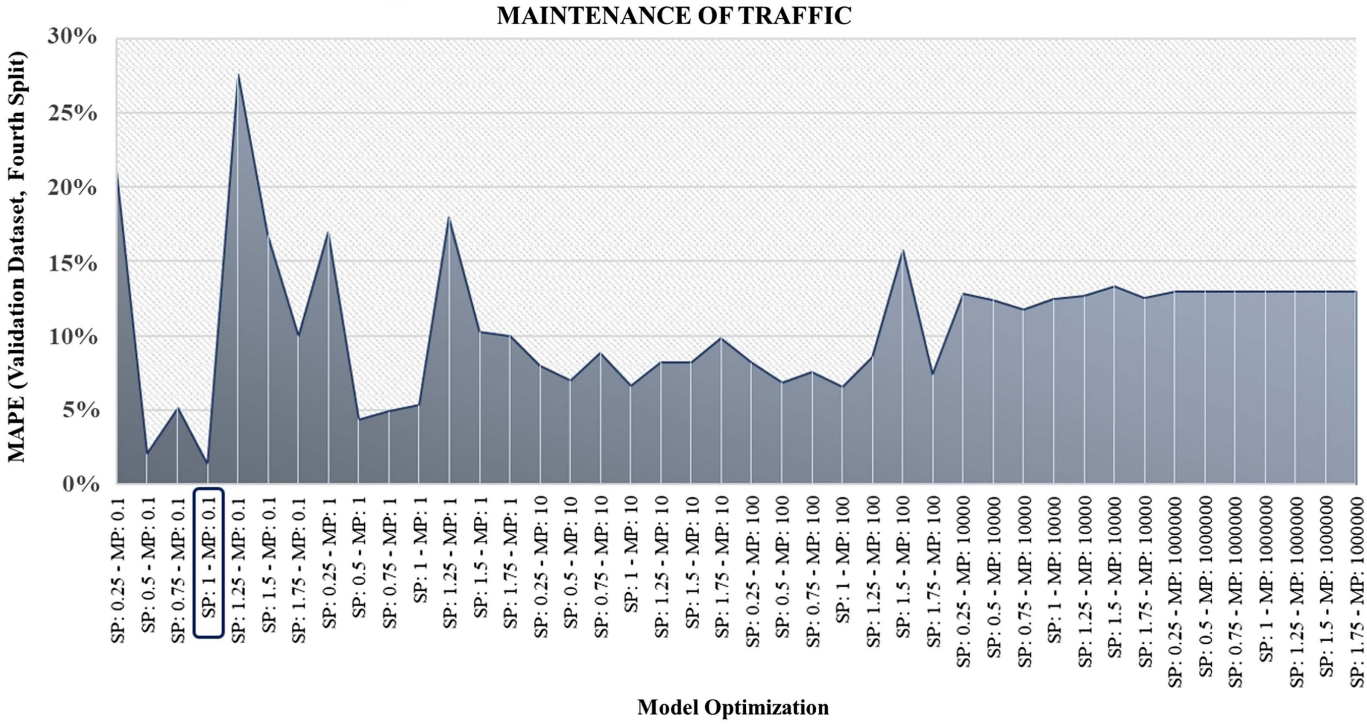

Fig. 11. Model optimization on the validation dataset for "MAINTENANCE OF TRAFFIC" on the fourth split.

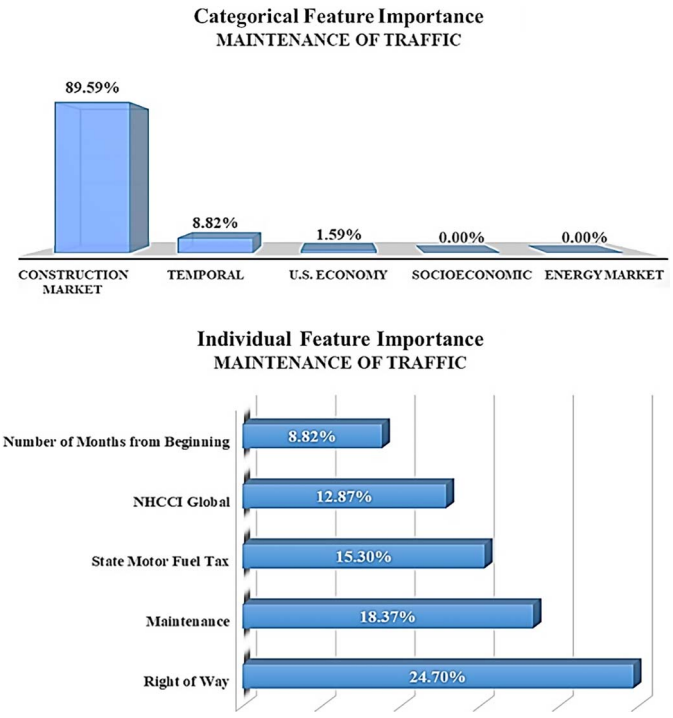

Fig. 12. Categorical and individual feature importance of "MAINTENANCE OF TRAFFIC."

demonstrate the application of the model and the level of accuracy. The "Results and Discussion" section of the study shows the features and models selected for the Florida dataset to show the application of the developed framework. All the selected features and models were data-driven, meaning that the authors did not select anything by choice or assumptions. The model (by comparing all the possible variations of variables and modeling algorithms) determined that the mentioned features have high importance in the cost prediction procedure for the Florida dataset. As a result, the complete framework is a comprehensive standalone approach that has much better generalization compared to the previous studies.

Table 4. Coverage and accuracy of the model

Construction type	Coverage of the total cost per mile of the model (%)	Prediction accuracy (%)
Constructing new urban 6L	93.97	90.95
Constructing new urban 4L	92.93	90.99
Constructing new rural 6L	98.27	93.62
Constructing new rural 4L	98.25	93.44
Widening 6L to 8L	86.84	93.00
Widening 4L to 6L	85.32	93.05
Average all construction types	92.60	92.51

Note: 4L = four lane; 6L = six lane; and 8L = eight lane.

The results of this research could also be used to attract private sector partnerships to foster economic development and improve safety and mobility. This would be accomplished by developing a suitable request for proposals and decent incentives accurately and on time by utilizing the proposed robust cost forecasting method. The main limitations of this research include data availability (50 cost items were available out of 60, and we only had access to 17 years of historical cost data) and the data level (using monthly level historical data). There are more complex modeling approaches such as deep learning, which were not included in the pipeline. The results showed that the chosen modeling approaches were capable of generalization in the case study. However, future studies based on the complexity of their specific case may need to

A1. Socioeconomic candidate variables				
#	Candidate variable	Acronyms	Scope	Source
1	Length Paved Roads (Centerline Miles)	LPRCMC	County	Florida Department of Transportation
2	Length Paved Roads (Line Miles)	LPRLMC	County	Florida Department of Transportation
3	Number of Household Estimates	HHEUS	County	U.S. Bureau of Census
4	Civilian Labor Force FL	CLFFL	Florida	U.S. Bureau of Labor Statistics
5	All Employees FL	AEFL	Florida	U.S. Bureau of Labor Statistics
6	Unemployment Rate	URUS	U.S.	U.S. Bureau of Labor Statistics
7	Change in Labor Market Conditions Index	CLMCIUS	U.S.	U.S. Bureau of Labor Statistics
8	Average Hourly Earnings Labor Employees: Construction	AHEPNECUS	U.S.	U.S. Bureau of Labor Statistics
9	Average Weekly Hours of All Employees Construction	AWHAECFL	Florida	U.S. Bureau of Labor Statistics
A2. Economy candidate variables				
#	Candidate variable	Acronyms	Scope	Source
1	Gross Domestic Products	GDP	U.S.	U.S. Bureau of Economic Analysis
2	Industrial Production	IP	U.S.	Federal Reserve System
3	Inflation Rate	IRUS	U.S.	World Bank
4	Consumer Price Index FL	CPIFL	Florida	U.S. Bureau of Labor Statistics
5	CPI for Urban Consumers: New vehicles	CPIAUCNV	U.S.	U.S. Bureau of Labor Statistics
6	CPI for All Urban Consumers: Used cars and trucks	CPIAUCUCT	U.S.	U.S. Bureau of Labor Statistics
7	Price Pressures Measure	PPMUS	U.S.	U.S. Bureau of Labor Statistics
8	Bank Prime Loan Rate	BPLRUS	U.S.	Federal Reserve System
9	30-Year Conventional Mortgage Rate	30YCMR	U.S.	Federal Reserve System
10	Leading Index for U.S.	LIUS	U.S.	Federal Reserve Bank
11	Leading Index for Florida	LIFL	Florida	Federal Reserve Bank
12	Producer Price Index for Commodities	PPIACO	U.S.	Federal Reserve Bank
13	Effective Federal Funds Rate	EDDRUS	U.S.	Federal Reserve Bank
14	M1	M1	U.S.	Federal Reserve Bank
15	M2	M2	U.S.	Federal Reserve Bank
16	Gold Prices	GP	U.S.	Yahoo Finance
17	Silver Prices	SP	U.S.	Yahoo Finance
18	Durable Goods Orders	DGOUS	U.S.	Yahoo Finance
19	Dow Jones Index Adj Close	DJI	U.S.	Yahoo Finance
20	S&P 500 Index	S&P500	U.S.	Yahoo Finance
21	St. Louis Fed Financial Stress Index	StLFFSI	U.S.	Yahoo Finance
22	Wilshire 5000 Total Market Full Cap	W5000TMFCI	U.S.	Yahoo Finance
23	NASDAQ Composite Index, Index	NASDAQ	U.S.	Yahoo Finance
24	Canada / U.S. Foreign Exchange Rate	CANUSER	U.S.	Yahoo Finance
25	China / U.S. Foreign Exchange Rate	CHUSER	U.S.	Yahoo Finance
26	Mexico / U.S. Foreign Exchange Rate	MEXUSER	U.S.	Yahoo Finance
27	U.S. / Euro Foreign Exchange Rate	USEUER	U.S.	Yahoo Finance
28	Total Vehicle Sales U.S.	TSVUS	U.S.	Yahoo Finance

A3. Construction market candidate variables				
#	Candidate variable	Acronyms	Scope	Source
1	New Private Housing Units (Building Permits)	NPHUABPFL	Florida	U.S. Bureau of Census
2	Construction Spending Nonresidential	CSNUS	U.S.	U.S. Census Burea
3	Construction Spending Highway	TCSHSUS	U.S.	U.S. Census Burea
4	Construction Employees	CEFL	Florida	U.S. Bureau of Labor Statistics
5	Construction Employees	AECHCEUS	U.S.	U.S. Bureau of Labor Statistics
6	NHCCI Global	NHCCI	U.S.	U.S. Bureau of Labor Statistics
7	Capital Expenditures	CEXFL	Florida	Florida State DOT Disbursements
8	DOT Disbursements: Maintenance	FDOTDM	Florida	Florida State DOT Disbursements
9	Administration Disbursements	ADFL	Florida	Florida State DOT Disbursements
10	Interest	IFL	Florida	Florida State DOT Disbursements
11	Bond Retirement	BRFL	Florida	Florida State DOT Disbursements
12	Local Government Grants	LGGFL	Florida	Florida State DOT Disbursements
13	Total Florida State DOT Disbursements	TFSDDFL	Florida	Florida State DOT Disbursements
14	Federal Funding	FFFL	Florida	Florida State Revenue Sources
15	State Motor Fuel Tax	SMFTFL	Florida	Florida State Revenue Sources
16	State Motor Vehicle Tax	SMVTFL	Florida	Florida State Revenue Sources
17	Tolls Income	TIFL	Florida	Florida State Revenue Sources
18	Other State Funding	OSFFL	Florida	Florida State Revenue Sources
19	Local Funding	LFFL	Florida	Florida State Revenue Sources
20	Total Florida State Revenue Sources	TFSRSFL	Florida	Florida State Revenue Sources
21	Right of Way	ROWFL	Florida	Florida Department of Transportation
22	Total State Highway System (SHS) PRODUCT AREAS	TSHSFL	Florida	Florida Department of Transportation
23	State Highway System (SHS) PRODUCT AREAS	SHSPAFL	Florida	Florida Department of Transportation
24	Other Roads PRODUCT AREAS	ORPAFL	Florida	Florida Department of Transportation
25	Legislative Budget Request	LBRAGFL	Florida	Florida Department of Transportation
A4. Energy market candidate variables				
1	Electricity Price	ELECFL	Florida	U.S. Energy Information Administration
2	Crude Oil Price	COP	U.S.	U.S. Energy Information Administration
3	Natural Gas Prices	NGP	U.S.	U.S. Energy Information Administration
4	Gas Price FL	GASPFL	Florida	U.S. Energy Information Administration

A5. Temporal candidate variables	
1	Number of Months from Beginning
2	Month
3	YEAR

Fig. 13. Independent variables employed in this study.

incorporate more complex modeling approaches. Finally, to enhance the level of the accuracy of the developed model, the next step would be to include the environmental, energy, and political trends as independent variables (predictors) in the pool of candidate variables. For future work, it is vital to investigate the impact of the contractors' management style on the highway construction cost. Ultimately, the same approach can be used for modeling and forecasting of highway maintenance cost. A complete construction and maintenance cost forecast platform can enable the users to see the life cycle cost of a highway in the planning stage, which would be an extremely valuable contribution.

Appendix I. Independent Variables (Predictors)

This study utilized five categories of independent variables (69 variables). The details of the predictors employed in this research are shown in Fig. 13.

Appendix II. Categorical Feature Importance Results

Cost item	Construction market (%)	Energy market (%)	Socioeconomic (%)	US economy (%)	Temporal (%)
MAINTENANCE OF TRAFFIC	89.59	0.00	0.00	1.59	8.82
MOBILIZATION	80.35	6.12	5.51	4.87	3.15
SEDIMENT BARRIER	59.51	0.00	7.10	26.38	7.01
FLOATING TURBIDITY BARRIER	69.18	0.00	15.73	12.63	2.47
STAKED TURBIDITY BARRIER-NYL REINF PVC	68.98	0.00	0.04	0.00	30.99
SOIL TRACKING PREVENTION DEVICE	92.64	0.00	5.91	1.45	0.00
LITTER REMOVAL	69.92	0.00	10.96	15.94	3.18
MOWING	82.63	0.00	6.33	8.67	2.37
CLEARING & GRUBBING	59.57	19.81	2.25	0.00	18.37
REGULAR EXCAVATION	80.54	0.00	9.65	8.96	0.85
EMBANKMENT	75.68	2.73	8.02	3.87	9.72
TYPE B STABILIZATION	73.51	0.32	11.97	14.10	0.09
OPTIONAL BASE, BASE GROUP 04	95.41	0.00	0.00	0.00	4.59
OPTIONAL BASE, BASE GROUP 09	75.59	0.00	2.77	21.64	0.00
MILLING EXIST ASPH PAVT, 1" AVG DEPTH	100.00	0.00	0.00	0.00	0.00
MILLING EXIST ASPH PAVT, 4" AVG DEPTH	82.54	0.00	6.63	7.92	2.90
SUPERPAVE ASPH CONC, TRAF C, PG76-22	83.38	0.00	4.98	0.74	10.90
SUPERPAVE ASPH CONC, TRAF D, PG76-22	81.21	2.66	7.11	3.10	5.92

Appendix II. (*Continued.*)

Cost item	Construction market (%)	Energy market (%)	Socioeconomic (%)	US economy (%)	Temporal (%)
ASPH CONC FC, INC BIT, FC-5, PG76-22, PMA	73.01	0.00	2.12	1.43	0.00
ASPH CONC FC, TRAFFIC B, FC-9.5, PG 76-22	62.53	0.00	0.00	17.80	1.95
CONC CLASS II, ENDWALLS	79.43	0.04	10.64	6.13	3.77
INLETS, DT BOT, TYPE D, <10′	73.98	5.98	8.19	8.25	3.61
INLETS, DT BOT, TYPE E, <10′	100.00	0.00	0.00	0.00	0.00
INLETS, BARRIER WALL, <= 10′	87.33	0.00	8.89	3.78	0.00
PIPE CULV, OPT MATL, ROUND,24"SD	61.02	0.00	18.41	18.27	2.30
PIPE CULV, OPT MATL, ROUND, 24"S/CD	56.17	0.00	19.99	18.41	5.44
PIPE CULV, OPT MATL, ROUND, 30"S/CD	75.43	0.70	16.13	7.46	0.29
PIPE CULV, OPT MATL, ROUND, 36"S/CD	52.93	1.22	22.37	17.59	5.89
PIPE CULV, OPT MATL, ROUND, 42"S/CD	76.78	0.00	2.54	0.00	20.68
PIPE CULV, OPT MATL, ROUND, 54"S/CD	79.22	2.20	7.13	2.37	9.09
MITERED END SECT, OPTIONAL RD, 24" SD	67.51	1.50	8.99	20.02	1.97
MEDIAN CONC BARRIER WALL	69.31	0.00	30.69	0.00	0.00
SHLDR CONC BARRIER, RIGID-SHLDR	89.56	0.00	9.97	0.47	0.00
CONCRETE DITCH PAVT, NR, 3"	99.70	0.00	0.00	0.30	0.00
RUMBLE STRIPS, GROUND-IN, 16" MIN. WIDTH	100.00	0.00	0.00	0.00	0.00
FENCING, TYPE B, 5.1-6.0', STANDARD	85.31	0.00	2.98	11.71	0.00
PERFORMANCE TURF	72.27	0.00	0.00	0.00	27.73
PERFORMANCE TURF, SOD	80.35	14.83	3.25	1.57	0.00
SINGLE POST SIGN, F&I GM, <12 SF	94.86	0.00	0.00	5.14	0.00
SINGLE POST SIGN, F&I GM, 12-20 SF	70.74	0.00	0.00	0.00	29.26
MULTIPOST SIGN, F&I GM, 21-30 SF	73.36	0.00	10.28	16.37	0.00
SINGLE POST SIGN, REMOVE	92.60	0.00	0.00	0.00	7.40
MULTIPOST SIGN, F&I GM, 31-50 SF	87.26	0.00	0.00	3.03	9.71
MULTIPOST SIGN, F&I GM, 51-100 SF	93.29	6.71	0.00	0.00	0.00
MULTIPOST SIGN, F&I GM, 101-200 SF	100.00	0.00	0.00	0.00	0.00
MULTIPOST SIGN, REMOVE	83.31	0.00	0.00	0.00	16.69
RETRO-REFLECTIVE/RAISED PAVEMENT MARKERS	45.28	0.00	23.00	7.51	24.20

Data Availability Statement

Data analyzed during the study were provided by a third party. Requests for data should be directed to the provider indicated in the Acknowledgements. Information about the *Journal*'s data-sharing policy can be found here: http://ascelibrary.org/doi/10.1061/(ASCE)CO.1943-7862.0001263.

Acknowledgments

The Florida Department of Transportation provided all the data needed for this project including the historical highway construction cost items. The researchers wish to thank the following FDOT individuals: Dianne Perkins, Cheri Sylvester, and June Mobley. We would also like to thank Hari Salkapuram (HDR Inc.), and Mansoor Khuwaja (Hanson Service Inc.).

References

Anderson, S., K. Molenaar, and C. Schexnayder. 2007. *Final report for NCHRP report guidance for cost estimation and management for highway projects during planning, programming and preconstruction*. Washington, DC: Transportation Research Board of the National Academics.

Emsley, M. W., D. J. Lowe, A. R. Duff, A. Harding, and A. Hickson. 2002. "Data modelling and the application of a neural network approach to the prediction of total construction costs." *Constr. Manage. Econ.* 20 (6): 465–472. https://doi.org/10.1080/01446190210151050.

Flyvbjerg, B., M. Skamris, and S. Buhl. 2002. "Underestimating costs in public works projects: Error or lie?" *J. Am. Plann. Assoc.* 68 (3): 279–295. https://doi.org/10.1080/01944360208976273.

Ji, S. H., and M. Park. 2010. "Data preprocessing-based parametric cost model for building projects: Case studies of North Korean construction projects." *J. Constr. Eng. Manage.* 136 (8): 844–853. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000197.

Kaliba, C., M. Muya, and K. Mumba. 2009. "Cost escalation and schedule delays in road construction projects in Zambia." *Int. J. Project Manage.* 27 (5): 522–531. https://doi.org/10.1016/j.ijproman.2008.07.003.

Kellerman, A. 2018. *Automated and autonomous spatial mobility*. Cheltenham, UK: Edward Elgar Publishing.

Kim, B. S., and T. Hong. 2012. "Revised case-based reasoning model development based on multiple regression analysis for railroad bridge construction." *J. Constr. Eng. Manage.* 138 (1): 154–162. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000393.

Lowe, D. J., M. W. Emsley, and A. Harding. 2006. "Predicting construction cost using multiple regression techniques." *J. Constr. Eng. Manage.* 132 (7): 750–758. https://doi.org/10.1061/(ASCE)0733-9364(2006)132:7(750).

Mahamid, I. 2011. "Early cost estimating for road construction projects using multiple regression techniques." *Constr. Econ. Build.* 11 (4): 87–101. https://doi.org/10.5130/AJCEB.v11i4.2195.

Membah, J., and E. Asa. 2015. "Estimating cost for transportation tunnel projects: A systematic literature review." *Int. J. Constr. Manage.* 15 (3): 196–218. https://doi.org/10.1080/15623599.2015.1067345.

Minchin, R. E., C. R. Glagola, K. Thakkar, and A. Santoso. 2004. "Managing preliminary estimate accuracy in a changing economy." In *Proc., American Society of Civil Engineers Specialty Conf. on Leadership and Management in Construction*, 120–128. Reston, VA: ASCE.

Pedregosa, F., et al. 2011. "Scikit-learn: Machine learning in Python." *J. Mach. Learn. Res.* 12 (2011): 2825–2830.

PWC (PricewaterhouseCoopers). 2016. *Public-private partnerships in the US: The state of the market and the road ahead*. London: PWC.

Python. 2020. "Python 3.8.3." Accessed January 12, 2020. https://www.python.org/downloads/.

Schach, R., and R. Naumann. 2007. "Comparison of high-speed transportation systems in special consideration of investment costs." *Transport* 22 (3): 139–147. https://doi.org/10.1080/16484142.2007.9638116.

Shahandashti, S. M., and B. Ashuri. 2016. "Highway construction cost forecasting using vector error correction models." *J. Manage. Eng.* 32 (2): 04015040. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000404.

Shojaei, A., and A. Mahdavian. 2019. "Revisiting systems and applications of artificial neural networks in construction engineering and management. Interdependence between structural engineering and construction management." In *Proc., 10th Int. Structural Engineering and Construction Conf.* Fargo, ND: International Structural Engineering & Construction Society.

Swei, O., J. Gregory, and R. Kirchain. 2017. "Construction cost estimation: A parametric approach for better estimates of expected cost and variation." *Transp. Res. Part B: Methodol.* 101 (Jul): 295–305. https://doi.org/10.1016/j.trb.2017.04.013.

Turochy, R. E., L. A. Hoel, and R. S. Doty. 2001. *Technical assistance report, highway project cost estimating methods used in the planning stage of project development*. Richmond, VA: Virginia DOT and the Univ. of Virginia.

Victoria Transport Policy Institute. 2016. *Transportation cost and benefit analysis techniques, estimates and implications*. 2nd ed. Victoria, BC: Victoria Transport Policy Institute.

Williams, T. P. 1994. "Predicting changes in construction cost indexes using neural networks." *J. Constr. Eng. Manage.* 120 (2): 306–320. https://doi.org/10.1061/(ASCE)0733-9364(1994)120:2(306).

Williams, T. P. 2003. "Modeling dredging project cost variations." *J. Waterway, Port, Coastal, Ocean Eng.* 129 (6): 279–285. https://doi.org/10.1061/(ASCE)0733-950X(2003)129:6(279).

Wilmot, C. G., and G. Cheng. 2003. "Estimating future highway construction costs." *J. Constr. Eng. Manage.* 129 (3): 272–279. https://doi.org/10.1061/(ASCE)0733-9364(2003)129:3(272).

Wilmot, C. G., and B. Mei. 2005. "Neural network modeling of highway construction costs." *J. Constr. Eng. Manage.* 131 (7): 765–771. https://doi.org/10.1061/(ASCE)0733-9364(2005)131:7(765).

Wong, J. M. W., and S. T. Ng. 2010. "Forecasting construction tender price index in Hong Kong using vector error correction model." *Constr. Manage. Econ.* 28 (12): 1255–1268. https://doi.org/10.1080/01446193.2010.487536.

Zhang, L. C., I. Johansen, and R. Nygaard. 2017. *Testing unit value data price indices*. Southampton, UK: Univ. of Southampton.