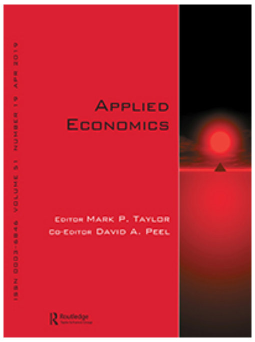

Applied Economics

Routledge
Taylor & Francis Group

ISSN: 0003-6846 (Print) 1466-4283 (Online) Journal homepage: https://www.tandfonline.com/loi/raec20

Predicting bid prices by using machine learning methods

Jong-Min Kim & Hojin Jung

To cite this article: Jong-Min Kim & Hojin Jung (2019) Predicting bid prices by using machine learning methods, Applied Economics, 51:19, 2011-2018, DOI: 10.1080/00036846.2018.1537477

To link to this article: https://doi.org/10.1080/00036846.2018.1537477

Published online: 30 Oct 2018.

APPLIED ECONOMICS
2019, VOL. 51, NO. 19, 2011–2018
https://doi.org/10.1080/00036846.2018.1537477

Predicting bid prices by using machine learning methods

Jong-Min Kim[a] and Hojin Jung[b]

[a]Statistics Discipline, Division of Science and Mathematics, University of Minnesota at Morris, Morris, MN, United States; [b]Department of Economics, Chonbuk National University, Korea

ABSTRACT

It is well-known that empirical analysis suffers from multicollinearity and high dimensionality. In particular, this is much more severe in an empirical study of itemized bids in highway procurement auctions. To overcome this obstacle, this article employs the regularized linear regression for the estimation of a more precise interval for project winning bids. The approach is put to the test using empirical data of highway procurement auctions in Vermont. In our empirical analysis, we first choose a set of crucial tasks that determine a bidder's bid amounts by using the random forest variable selection method. Given the selected tasks, project bid forecasting is conducted. We compare our proposed methodology with the least square linear model based on the bias and the standard root mean square error of the bid estimates. There is evidence supporting that the suggested approach provides superior forecasts for an interval of winning bids over the competing model. As far as we know, this article is the first attempt to provide reference bids of highway construction contracts.

KEYWORDS
Auctions; bidding; regularized linear regression; forecasting

JEL CLASSIFICATION
C53; D44; H57

I. Introduction

Public procurement accounts for roughly 10–15% of GDP for developed countries and can amount to as much as 20% of GDP for developing countries (Kashap 2004). The public sector spends between $1.4 and $1.6 trillion annually. In particular, the United States federal government alone spent $231.08 billion in 2000 while state and local governments spent about six times more than the federal government in the procurement process (Thai 2001). In the state of Vermont, the state government spent about 8% of its total spending in transportation projects.[1] Approximately $0.2 billion of this budget was spent on road construction alone. The auction mechanism is typically used for public procurement, especially highway construction contracts. However, there have only been limited studies on bidding function estimation, which take into account itemized bids due to the problem of high-dimensionality. Most previous studies on bidding function estimation consider overall project bids instead of itemized bids (see, Krasnokutskaya and Seim 2011; Marion 2007; De Silva, Dunne, and Kosmopoulou 2003; Jofre-Bonet and Pesendorfer 2000).

The purpose of this article is to adopt an empirical method to estimate a reliable interval for project bids as reference winning bids in practice. While a few previous studies (e.g. Bajari, Houghton, and Tadelis 2014; Jung et al. 2016) attempt to estimate the bidding functions by using itemized bids, they fail to provide a complete analysis due to the 'curse of dimensionality'. Their model specifications are likely to be misspecified. Therefore, such a bidding function estimation provides a limited forecast of the winning bids. Forecasting ability is important because it is readily relevant for potential bidders to predict the possible winning bids when they determine their bid amounts. It also helps state agencies predict what their construction budgets are going to be and determine whether to re-let a project if the received bids do not fit within predictions.

In this article, the interval of winning bids is estimated based on tasks frequently used for highway construction. After selecting key tasks by using the random forest (RF) method, the regularized

CONTACT Hojin Jung hojin.jung@jbnu.ac.kr Chonbuk National University, Jeonju-si, Jeollabuk-do 54896, Korea

[1]Source: usgovernmentspending.com.

linear regression method can be used to predict the intervals of winning bids. Since targeting attributes of each highway construction differ distinctly, direct prediction of each project is not effective. Therefore, our forecasting models are used to predict the bid distribution containing a winning bid. This functionality is very useful for bidders in practice to obtain bid guidance.

Many state governments provide a detailed project description, including state engineer's cost estimates, which reduce informational asymmetries among bidders (De Silva, Kosmopoulou, and Lamarche 2009). It includes estimates of quantities and prices of each task in the project. The effect of sharing information on bidding behavior has been considered since the early theoretical paper by Milgrom and Weber (1982), in which they prove that releasing information causes rational bidders to shade their bids to offset the winner's curse in common value auction. Goeree and Offerman (2002) also show that releasing information helps bidders reduce uncertainty about the value of an item and the winner's curse in their experimental setting. De Silva et al. (2008) support this theory using data from road construction in Oklahoma and Texas. Furthermore, De Silva, Kosmopoulou, and Lamarche (2009) find that the release of information will help new entrants survive longer in the construction market.

Previous literature has developed sophisticated techniques for estimating bidders' bidding functions. For example, Guerre, Perrigne, and Vuong (2000) propose a two step procedure for nonparametrically estimating distribution of bidders' private values from observed bids in a symmetric independent private value framework. Goswami and Wettstein (2016) analyze bidding behavior in a subjective environment, describing a class of auctions where the bidders do not precisely know how their bids will be evaluated, by introducing subjectivity. Fry et al. (2016) explore the impact of better cost estimation accuracy for a firm that bids on projects. Their results suggest that firms with more accurate cost estimation are more likely to lower their bid amounts and thus are more likely to win more projects than firms that have less cost estimation accuracy.

The layout of the paper is as follows. Section 2 is devoted to describing our data and general procurement auction procedures by the Vermont Agency of Transportation (VTrans). The RF variable selection method and the regularized linear regression approach are discussed in Section 3. In Section 4, we discuss empirical analysis. Concluding remarks are presented in Section 5.

II. Highway construction auction

In this analysis, we use information on road construction projects procured by the VTrans from 2006 through 2009. While there are other types of projects auctioned off during the sample period, this article focuses on highway projects only. The state auctions off projects on a weekly basis and uses sealed-bid tenders, where the contract is awarded to the lowest qualified bidder. The VTrans advertises upcoming auctions for about 3 weeks and provides brief information, such as the project work site and estimates of the number of working days for the project. A firm in Vermont is assigned a level of qualification of bidding that determines the value of the projects and the number of contracts that it can undertake within a year. Each firm's pre-qualified status is determined based on the firm's current financial status and performance history in the Vermont highway construction industry. After the contract is awarded, the identities and the bids of all bidders are made available to the public. Therefore, we have information on the bids of all bidders and the winner for each project.

Figure 1 illustrates the geographic distribution of the projects. The blue circle indicates the base project used for selecting important tasks in our empirical analysis. The red marks show the sample contracts used for evaluating the predictive ability of our proposed method in forecasting.[2] As shown in the figure, most sample projects are located near interstate highways (I-89 and I-91) in Vermont. Note that the selected sample contracts also include those in remote/less populated areas or in mountainous terrain that might affect project uncertainty, thus project bids.

[2] A triangle, a square, a pentagon, a hexagon, an octagon, and a round square indicate the sample projects of 04B208, 04C138, 04C178, 05B126, 04B198, and 06B240 respectively.

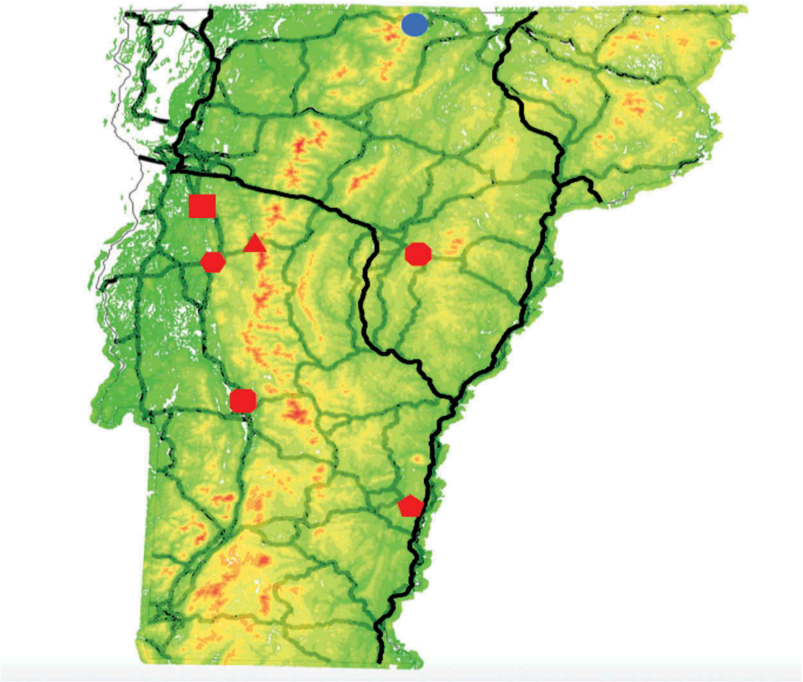

Figure 1. Project distribution.

III. Econometric methodology

Random forest variable selection method

In this section, we select the most important tasks that determine bidders' bid amounts by using one of the most popular machine learning methods, the RF method. This method is an ensemble classification algorithm, which uses trees as base classifiers. It performs well on classification and regression with many different types of datasets (Breiman 2001; Hastie, Tibshirani, and Friedman 2009). RF is an ensemble approach, which is based on the bootstrap and selection of random subset of predictor variables as candidates for splitting tree nodes. This method constructs many decision trees that individual learners combined. It is also known that RF produces good out-of-sample fits for highly nonliner data (Varian 2014). The basic idea of RF is to train an ensemble of uncorrelated, weak learners with high variance on bootstrap samples of the data, and then average the resulting output. See James et al. (2013) for more detailed descriptions of various methods for modeling complex datasets. In this study, we implement Breiman's RF algorithm for regression via the R package *randomForest*. The RF regression prediction for a new observation x $(\hat{f}_{rf}^{B}(x))$ is made by averaging the output of the ensemble of B trees $\{T(x, \Psi_b)\}_1^B$ as

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T(x, \Psi_b),$$

where Ψ_b characterizes the b^{th} RF tree in terms of split variables, cutpoints at each node, and terminal node values.

RF provides an approach to select feature ranking, such as mean decrease impurity and mean decrease accuracy. It can also be used in unsupervised mode for assessing proximities among data points. Figure 2 displays errors versus the number of trees obtained by RF. We make a RF of 70 trees, and importance of predictors is assessed to minimize prediction error on the training set. The other parameters are kept at their default settings. RF also provides a variable importance level which is measured by the expected fraction of the splits that a feature contributes to. In other words, the number of times a feature is used to provide an optimal split in a node of a decision tree in the forest can provide an estimate of the relative importance of a feature.

Feature importance in RF is measured by randomly permuting the feature in the out-of-bag (OOB) samples and calculating the percent increase in misclassification rate as compared to the OOB rate with all variables intact. The OOB samples are used to measure the prediction accuracy in the model. For

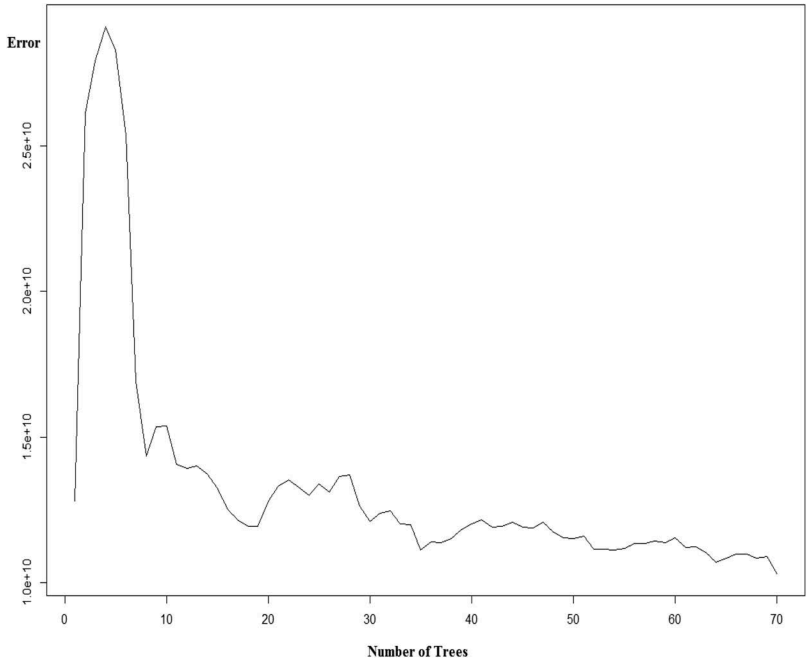

Figure 2. Errors plot.

regression in particular, the prediction error on the OOB portion is used to calculate the mean square error. We select the important variables based on the mean decrease in node impurities from splitting on the variable shown on the right panel of Figure 3. The node impurity is measured by residual sum of squares. The important variables chosen by RF from our base project of 04B140 (2006) in this article are the items of 204.30, 404.65, 540.10, 613.11, 621.215, 621.50, 631.17, 651.18, 651.20, 651.40, and 900.645.[3]

Regularized linear regression

It is not unusual to see a high-dimensional case in which the number of observed independent variables greatly exceeds the number of observations. In such a case, without penalization one could fail to obtain a least squares (LS) regression estimator or appropriate prediction intervals. Another possible issue regarding our sample dataset could be that $X'X$ could be singular or nearly singular, which means that explanatory variables are highly correlated. In this scenario, the LS estimates are unbiased, but their variances could be much larger. The regularized linear regression could overcome these problems by assigning a penalty term to reduce the standard errors.

We consider two regularized linear regression methods: Ridge and least absolute shrinkage and selection operator (Lasso). The Ridge method is useful when some of true parameters are zero or close to zero. The underlying assumptions in this regression are the same as LS, with the exception of the normality assumption in LS. Given the response vector $y \in R^n$ and the measurement matrix $X \in R^{n\times p}$, the estimate of β is obtained by minimizing the following Ridge constraint:

$$\underset{\beta\in R^n}{\operatorname{argmin}}\ (y - X\beta)'(y - X\beta) \quad \text{subject to}$$

$$\sum_{j=1}^{p} \beta_j^2 \leq t.$$

The Ridge constraint can be expressed as the following penalized residual sum of squares (PRSS)

$$\text{PRSS}(\beta) = (y - X\beta)'(y - X\beta) + \lambda||\beta||_2^2,$$

where λ is a shrinkage parameter, which controls the amount of regularization and thus, the size of

[3] The pay item description for the items listed is the following: 204.30: Granular Backfill for Structures Bituminous Concrete Pavement, 404.65: Emulsified Asphalt, 540.10: Placing Concrete, 613.11: Stone Fill, Type II, 621.215: HD Steel Beam Guardrail, Galvanized w/2.4 m (8 feet) Posts, 621.50: Manufactured Terminal Section, Flared, 631.17: Testing Equipment, Bituminous, 651.18: Fertilizer, 651.20: Agricultural Limestone, 651.40: Grubbing Material, and 900.645: Special Provision.

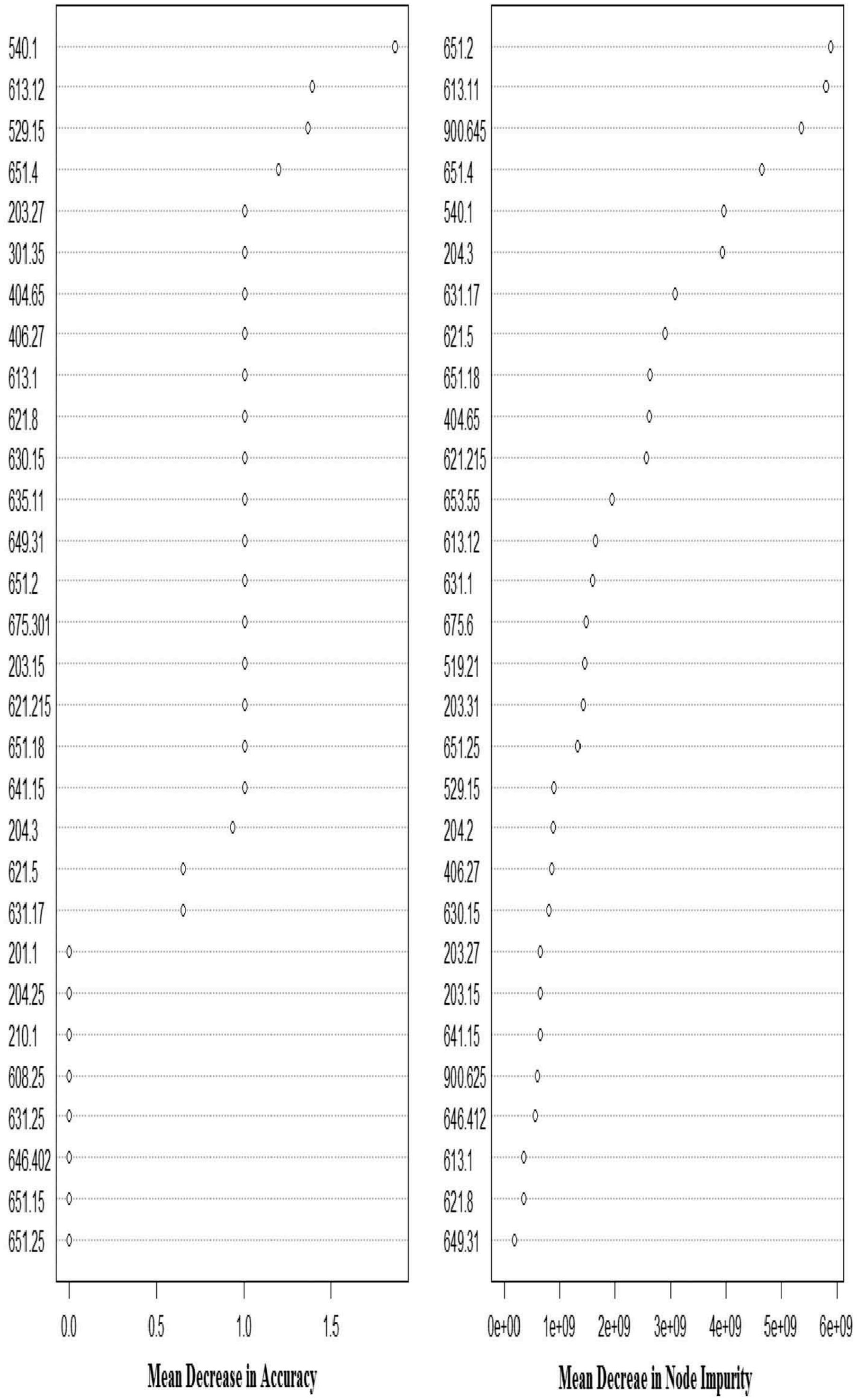

Figure 3. Variance important plot.

the coefficients. The optimization solution is obtained by taking derivatives:

$$\frac{\partial \mathrm{PRSS}(\beta)}{\partial \beta} = -2X'(y - X\beta) + 2\lambda\beta$$

$$\hat{\beta}^{Ridge} = \left(X'X + \lambda I_p\right)^{-1} X'y.$$

Similar to the Ridge regression, the Lasso, which was introduced by Tibshirani (1996), is capable improving the accuracy of LS models. The estimate of β is the solution to the following optimization problem

$$\underset{\beta \in R^n}{\operatorname{argmin}} \ (y - X\beta)'(y - X\beta) \quad \text{subject to}$$
$$\sum_{j=1}^{p} |\beta_j| \le t.$$

Equivalently, the objective function can be expressed as the following loss function:

$$\mathrm{PRSS}\,(\beta) = (y - X\beta)'(y - X\beta) + \lambda ||\beta||_1.$$

Unlike Ridge regression, the Lasso has no closed form solution. Notice that the only difference between the two methods is the penalty function. Also these two methods will be the same as LS if $\lambda = 0$. We use R package *lars* in our empirical analysis.

IV. Illustrated example

We use the bid history of the highway construction contracts auctioned off in Vermont in our empirical analysis.[4] Among important variables selected by RF from the base contract in Section 3, we further restrict our attention to the five common variables (204.30, 404.65, 631.17, 651.18, and 651.20) from our sample contracts.[5] We empirically model the firm's bid as a linear function of the restricted tasks. It is not surprising that every task selected by RF is statistically significant. In particular, R^2 is 0.999, indicating that our control variables explain almost 99.9% of the variations of the firm's total bids. Thus we are confident that our empirical analysis is conducted using the appropriate explanatory variables.

We carry out empirical exercises to estimate a bidder's bid by using the proposed method. To evaluate the forecasting power of the method, we use two statistics, such as the bias and the standard root mean square error (RMSE) of the bid estimates. Table 1 presents the results from the model comparisons. From the empirical bias defined as a firm's project bids minus its estimate, it is evident that our methodology provides at least as accurate forecasts as a simple linear LS model. MSE is obtained via the following formula: $\frac{1}{n}\sum_{i=1}^{n}(b_i - \hat{b}_i)^2$, where n is the number of bidders and b is the observed bid in an auction. RMSE is the square root of the MSE. A value is highlighted in the table if the bias in absolute value or the RMSE is less than that of the linear LS model. It is worth mentioning that the LS model shows better accuracy for some projects in the sample. However, the difference in the accuracy between the LS model and each regularized linear regression is between $105 and $498 which is a negligible amount relative to its overall project values.

According to the attributes that were described in Section 3, the task now is to predict the interval for the winning bids of a highway construction project. The results in Table 2 show the estimated

Table 1. Absolute Bias and RMSE.

	Method	Bidding Project Number					
		04B208	04C138	04C178	05B126	04B198	06B240
Absolute Bias	Ridge	66.415	**169,636**	435.281	**275.955**	654.939	**143.807**
	Lasso	66.415	**169,636**	435.281	**275.955**	654.939	**143.807**
	LS	**66.311**	169.789	**434.783**	276.350	**654.724**	144.201
	Difference (a)	0.105	−0.153	0.498	−0.395	0.215	−0.393
	Difference (b)	0.105	−0.153	0.498	−0.395	0.215	−0.393
RMSE	Ridge	101.325	**198.302**	456.291	**337.847**	701.249	**164.125**
	Lasso	101.325	**198.302**	**455.875**	338.304	**701.062**	164.530
	LS	**101.177**	198.425	**455.875**	338.304	**701.062**	164.530
	Difference (c)	0.148	−0.123	0.416	−0.457	0.187	−0.405
	Difference (d)	0.148	−0.123	0.000	0.000	0.000	0.000

All monetary figures are expressed in 1,000 dollars. Difference (a) = $\text{bias}^{Ridge} - \text{bias}^{LS}$; Difference (b) = $\text{bias}^{Lasso} - \text{bias}^{LS}$; Difference (c) = $\text{RMSE}^{Ridge} - \text{RMSE}^{LS}$; Difference (d) = $\text{RMSE}^{Lasso} - \text{RMSE}^{LS}$

[4] The dataset of this article is publicly available from the Vtrans website.

[5] Note that these items frequently occur on a highway construction contract in practice. According to the speck book published by the VTrans, item 204.30 is one of the tasks for excavation and backfill for structures. Item 404.65 is used for furnishing and applying bituminous treatment. Item 651.18 is for erecting, equipping, and maintaining field offices and testing equipment. The last two items consist of the preparation of the area and the application of topsoil, grubbing material, sod, seed, soil amendments, and mulch.

Table 2. Min/Max for winning bidding estimation.

	Bidding Project Number					
Method	04B208	04C138	04C178	05B126	04B198	06B240
Lasso	[**193.593**, 268.758]	[**562.226**, 926.730]	[590.624, **1,189.753**]	[**357.974**, 1,234.263]	[296.798, **1,188.606**]	[**460.179**, 880.090]
Ridge	[**193.593**, 268.758]	[**562.226**, 926.730]	[590.624, **1,189.753**]	[**357.974**, 1,234.263]	[296.798, **1,188.606**]	[**460.179**, 880.090]
LS	[193.847, 268.800]	[562.667, 926.622]	[590.972, 1,190.592]	[590.972, 1,190.592]	[296.816, 1,188.802]	[460.449, 880.662]
Winning bid	243.624	504.892	1,065.682	342.908	1,254.924	427.791
Number of bidders	3	4	8	11	9	6

All monetary figures are expressed in 1,000 dollars.

intervals for predicting winning bids. The lower (upper) bound is the minimum (maximum) value of the estimated bids in an auction. The intervals are compared to the actual winning bids. Table 2 clearly presents a systematic pattern in the sense that the lower bound of the estimated interval is very close to the actual winning bid for a small-sized project while the upper bound is similar to the winning bid for a large-sized project. Given this pattern, Ridge and Lasso are significantly superior to the linear LS model. Our proposed method could provide a better way to predict specific winning bid ranges given project tasks in a highway procurement auction. Thus a bidder could have higher a probability of winning.

It is important to point out that our proposed method could have a stronger predictive power if the base project used for selecting the important tasks has similar auction specific characteristics to a project for estimation. In practice, a bidder could identify a similar base project based on the level of complexity and size for the project by using a prior way. For instance, previous literature proxies the number of different items for project complexity. It is also noteworthy that our empirical analysis is based on ex post bid amounts. However, that will not cause bias and inconsistency due to the stable prices of tasks over our sample periods. Our suggested approach is a generalized form that provides different key tasks in each auction because the potential bidder cohorts are different. Thus bidders will be able to obtain the intervals for the possible winning bids constructed by the information on the bidding history for all plan-holders. In the state of Vermont, firms will have information of their potential competitors in an auction when they prepare their bids. The Vtrans releases the information on plan-holders' identities if the number of plan-holders is greater than 3 in an auction. The number of bidders is on average less than 3.5 in our sample dataset, possibly due to the fact that Vermont is a relatively small sized state. Our intuition is that the predictive power of our proposed model could be improved by using a dataset of the states with large auction participants such as Texas, California, or New York.

V. Conclusion

This article introduced a strategy framework for successful bidding. We provided forecasting models for a more precise interval of winning bids. A typical project consists of 40 different tasks in our sample. There could be two important issues raised by our sample data. Empirical analysis will suffer from the multicollinearity and the high dimensionality. Another issue is that a traditional estimation will not be feasible due to limited data. To effectively deal with these potential problems, we selected tasks that are frequently used for highway construction by using the RF method. After selecting key tasks, statistical methods were used to predict the interval for winning bids in the highway contract auctions.

In particular, the forecasting was conducted by using the well-known regularized linear regression approaches: Ridge and Lasso. In order to evaluate our proposed method, we compared our estimated interval with the observed winning bids from our sample data. We found that the method is very useful for bidders to obtain bid guidance in practice. In addition, this approach would be most useful to state agencies who are trying to predict what their construction budgets are going to be and to help determine whether to re-let a project if the received bids do not fit within predictions.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Bajari, P., S. Houghton, and S. Tadelis. 2014. "Bidding for Incomplete Contracts: An Empirical Analysis of Adaptation Costs." *The American Economic Review* 104 (4): 1288–1319.

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

De Silva, D., T. Dunne, A. Kankanamge, and G. Kosmopoulou. 2008. "The Impact of Public Information on Bidding in Highway Procurement Auctions." *European Economic Review* 52 (1): 150–181.

De Silva, D., G. Kosmopoulou, and C. Lamarche. 2009. "The Effect of Information on the Bidding and Survival of Entrants in Procurement Auctions." *Journal of Public Economics* 93 (1): 56–72.

De Silva, D. G., T. Dunne, and G. Kosmopoulou. 2003. "An Empirical Analysis of Entrant and Incumbent Bidding in Road Construction Auctions." *The Journal of Industrial Economics* 51 (3): 295–316.

Fry, T. D., R. A. Leitch, P. R. Philipoom, and Y. Tian. 2016. "Empirical Analysis of Cost Estimation Accuracy in Procurement Auctions." *International Journal of Business and Management* 11 (3): 1–10.

Goeree, J. K., and T. Offerman. 2002. "Efficiency in Auctions with Private and Common Values: An Experimental Study." *The American Economic Review* 92 (3): 625–643.

Goswami, M. P., and D. Wettstein. 2016. "Rational Bidding in a Procurement Auction with Subjective Evaluations." *International Journal of Industrial Organization* 44: 60–67.

Guerre, E., I. Perrigne, and Q. Vuong. 2000. "Optimal Nonparametric Estimation of First-Price Auctions." *Econometrica* 68 (3): 525–574.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning*. New York. USA: Springer.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An Introduction to Statistical Learning*. Vol. 6. New York: springer.

Jofre-Bonet, M., and M. Pesendorfer. 2000. "Bidding Behavior in A Repeated Procurement Auction: A Summary." *European Economic Review* 44 (4): 1006–1020.

Jung, H., G. Kosmopoulou, C. Lamarche, and R. Sicotte (2016). Strategic Bidding and Contract Renegotiation. Working Paper.

Kashap, S. 2004. "Public Procurement as a Social." *Economic and Political Policy* 3: 133–147.

Krasnokutskaya, E., and K. Seim. 2011. "Bid Preference Programs and Participation in Highway Procurement Auctions." *The American Economic Review* 101 (6): 2653–2686.

Marion, J. 2007. "Are Bid Preferences Benign? the Effect of Small Business Subsidies in Highway Procurement Auctions." *Journal of Public Economics* 91 (7): 1591–1624.

Milgrom, P. R., and R. J. Weber. 1982. "A Theory of Auctions and Competitive Bidding." *Econometrica* 50 (5): 1089–1122.

Thai, K. V. 2001. "Public Procurement Re-Examined." *Journal of Public Procurement* 1 (1): 9–50.

Tibshirani, R. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of Royal Statistical Society, Series B* 58 (1): 267–288.

Varian, H. R. 2014. "Big Data: New Tricks for Econometrics." *The Journal of Economic Perspectives* 28 (2): 3–27.