Predicting Award Prices of First Price Sealed Bid Procurement Auctions

Fabian Blasch

Supervisor: Dr. Katharina Fenz

August 2, 2022

Contents

1 Introduction

The importance of public procurement becomes quickly apparent when looking at the sheer volume of contracts that is awarded through public auctions. The authorities of the European Union for example spent around 14% of their GDP on public procurement in 2017 (Rodríguez et al., 2020). Similar observations can also be made for many states in the U.S. One example of particular importance for this thesis is Colorado. The Colorado Department of Transportation (CDOT), is responsible for procurement of street and bridge construction contracts. As displayed below the budget for transportation is ranked as number four on the largest capital expenditures in the state's budget, after education, health care and human services. Of the approximate 2 billion dollars spent on transportation in 2021, the CDOT awarded $790 million in contracts, to design, repair and create bridges and highways (CDOT, 2021).

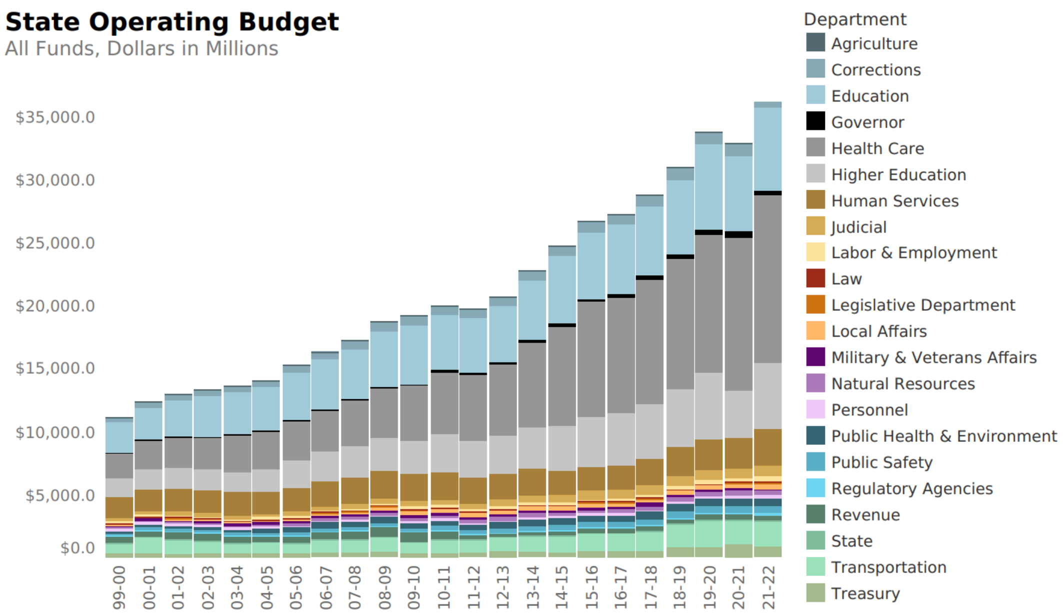

Figure 1: State of Colorado Budget

Given that the contracts awarded through procurement auctions are such a substantial part of the budget, every potential improvement to the process may greatly increase the efficiency of the tax payers money that is spent. Accordingly, a closer examination of the process including a prediction model for the auctions' award prices may support public procurement agencies like the CDOT with budget planning. Additionally, further analysis of the underlying data in respect to the interactions of particular bidders and the associated effect on award prices is also highly relevant for public procurement agencies. This thesis thus provides an analysis of different models, that predict the award price of an auction given input information available through the bid tabs that are published on the official website of the CDOT. In particular, four different model types with varying preprocessing schedules are compared in terms of their predictive power. Standard linear regression, elastic net regression, random forests and an eXtreme gradient boosting model. To assess, whether, the combination of certain bidders leads to higher award prices, recently discovered post-selection inference methods are utilized (Tibshirani et al., 2019)

The remaining paper is structured as follows, first the data extraction process from the PDFs provided on the CDOT's website is described. Then the general process of procurement is outlined, this also entails a concise description of the auction process utilized by the CDOT. The following section covers the methods that are applied, not only in respect to the different models used for prediction but also for the different preprocessing schedules that are applied and the post-selection inference that is used. The thesis concludes with the results for the best predictive model utilizing linear and quadratic loss functions and the results of the analysis of bidder interactions.

2 Data

All the information about the procurement contracts, is obtainable through the bid tab archive on the official website of the Colorado Department of Transportation. The information is provided in PDF documents. In each of those files the following information of the respective auction is provided.

- A table listing all submitted bids, including a unique identifier for each of the participating bidders
- A contract description
- An engineer's estimate
- The contract ID
- The letting date
- Either the amount of time given to complete all the contractual obligations, or a completion due date
- The county in which the contract is to be completed in

For illustrative purposes, Figure 2 displays an example of a bid tab, in particular the second page, which contains the vendor ranking as well as the contract description and the remaining information listed above.

Colorado Department Of Transportation Printed On: **11/17/2015**
Vendor Ranking Page 1 of 1

Letting No: 20151112
Letting Date: November 12, 2015
Letting Time: 10:00 AM

Contract ID: C19868
Region: 1
Contract Time: 260 WORKING DAYS

Project(s): STU1211-084
Counties: JEFFERSON, REGION 1

Contract Description:
SH121(WADSWORTH)-HIGHLAND DR-10TH AVE-JEFFERSON CO

THIS PROJECT IS LOCATED ON WADSWORTH BETWEEN HIGHLAND AND 10TH.

CONSTRUCTION WILL INCLUDE A FULL CONSTRUCTION WITH WIDENING OF ONE LANE IN BOTH DIRECTIONS, AND A MULTI MODAL TRAIL ON BOTH SIDES. THE MAINLINE PAVING WILL BE CONCRETE. THE WORK ALSO INCLUDES A CONCRETE BOX CULVERT NEAR HIGHLAND TO CARRY LAKEWOOD GULCH UNDER WADSWORTH.

CDOT WILL ONLY BE ACCEPTING ELECTRONIC BIDS FOR THIS PROJECT. PLEASE CONTACT BID EXPRESS CUSTOMER SERVICE AT 1-888-352-2439 TO OBTAIN AN ACCOUNT IF NECESSARY.

Rank	Vendor ID	Vendor Name	Total Bid	Percent Of Low Bid	Percent Of Estimate
0	-EST-	Engineer's Estimate	$9,821,027.20	91.58%	100.00%
1	870A	SEMA CONSTRUCTION, INC.	$10,723,550.00	100.00%	109.19%
2	884A	HAMON INFRASTRUCTURE, INC.	$10,817,000.00	100.87%	110.14%
3	1275A	CASTLE ROCK CONSTRUCTION COMPANY OF COLORADO, LLC	$10,817,845.03	100.88%	110.15%
4	065A	CONCRETE WORKS OF COLORADO INCORPORATED	$11,614,565.78	108.31%	118.26%
5	232A	AMERICAN CIVIL CONSTRUCTORS, INC. dba ACC Mountain West	$12,338,888.00	115.06%	125.64%

Figure 2: Bid Tab Example

2.1 Scraping

In order to obtain all the archived bid tabs, the html code of the website was first examined using a google chrome extension called SelectorGadget. This tool allows one to identify html nodes, that website contents are associated with. In the case of the bid tab archive, the html node carrying the links to the individual bid tabs is "<td a>". Once this html node is discovered and the consistency across different years in the archive is ensured, the download is easily achieved by looping over the links and downloading the respective PDFs. The hyperlink extraction was performed utilizing *rvest*, by Wickham (2022). For the remaining steps in the data extraction process, a distinction will be made for text based information and tabular data.

2.1.1 Text Based Information

The structure of the text based information allows us to filter the individual parts via regular expressions. Especially, for the letting data, the contract ID, and the county this required no further data cleaning steps. Unfortunately, this is not the case for the contract time and the contract description.

The contract time was not as straightforward to obtain, since the way it is reported is inconsistent across documents. Most of the time, it is reported as working days until all contractual obligations have to be fulfilled. Seldom, however, the bid tab contains a completion date instead. Accordingly, to achieve consistency across documents all completion dates were converted to contract time. This was achieved by first adding 60 days to the letting date, as this is the number of days that the CDOT reports as the expected time between the letting date and the start of the work on site. Then, the difference in days between the completion date and the starting date were computed. As said difference is only supposed to contain working days the following holidays as well as all weekends were subtracted from the difference between starting date and completion date.

- New Year's Day
- Dr. Martin Luther King, Jr. Day
- President's Day
- Memorial Day
- Juneteenth
- Independence Day
- Labor Day
- Frances Xavier Cabrini Day
- Veterans Day
- Thanksgiving
- Christmas

The computation was executed utilizing the R package *bizdays*, Freitas (2022). The package enables the user to generate custom calenders. The difference in starting and completion date was therefore easily calculated by setting up a custom calender with the holidays listed above as well as all Saturdays and Sundays. Then using this calender, the difference between two dates only takes working days into account. The only remaining text based information is the contract description. So far, none of the text based information required extensive preprocessing to obtain variables that can be represented in a tabular format. In the case of the contract description this is not the case. In order to convert the contract description into a format that may be represented in a table, the descriptions were first tokenized. Tokenization refers to splitting the input text into single unique words, i.e., splitting the sentences on spaces and removing all forms of punctuation. The result is then a vector of tokens. Said tokens were then scanned for spelling mistakes utilizing the R package *hunspell* (Ooms, 2020). Once the misspelled words were corrected, stop words were removed from the list of tokens. Stop words are words that have no inherent signal associated with their use, examples for such words in the English language would be “a”, “is” and “the”. In natural language processing there is not necessarily one list of stop words, depending on the context different libraries of stop words may be used to remove as much noise as possible from textual data while leaving the signal associated with a series of words in tact. In the case of this thesis, a combination of five different libraries of stop words was used. All of those libraries, “snowball” , “stopwords-iso”, “smart”, “marimo” and “nltk” are available through the R package *stopwords*, by Benoit et al. (2021). After filtering out the stop words, the remaining words were then stemmed. Stemming refers to the process in which a word is reduced to it’s root. This means, that words that carry an identical signal are reduced to the same shortest common substring. Consider the following three words, replacing, replaced, replacement. All those words carry the information that something needs replacement. The language specific circumstances that determine the affixes are not relevant for the information extraction and thus all the aforementioned words are shortened to “replac” Silge and Robinson (2017). After stemming, to remove any remaining misspelled words and also for potential removal of unwanted information, all stemmed words were written to an excel file and checked manually. Given that all stop words were already removed and the remaining words were reduced to their stem, this was a very feasible task, resulting in a file with around 2000 words to check. Below we observe the top 40 most frequent words that result from our text mining endeavours.

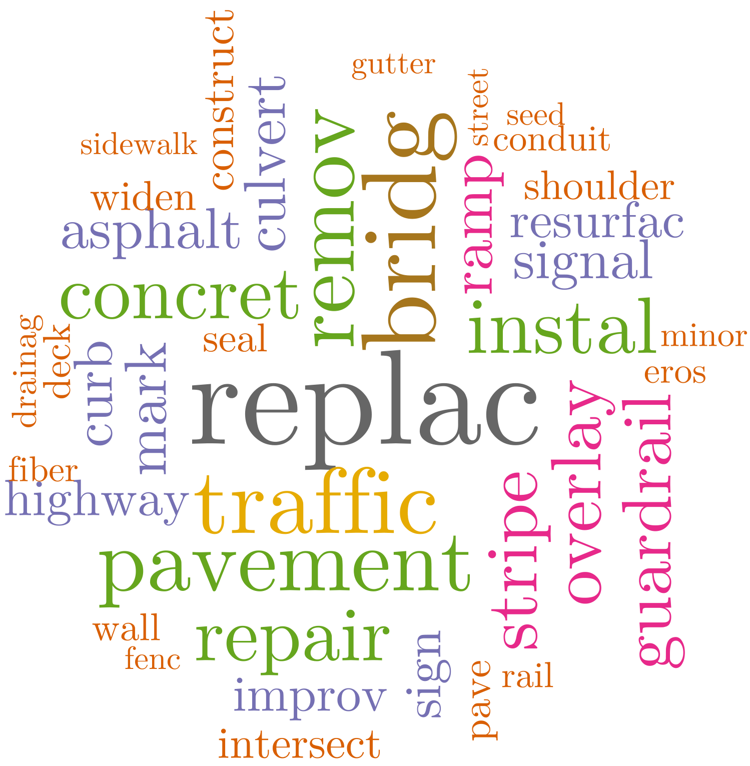

Figure 3: Top 40 Stemmed Description Words

2.1.2 Tabular Data

As displayed in Figure 2, the table in each of the auctions' PDFs contains information on the submitted bids, the bidders' identity and an engineer's estimate. To extract the table containing this information, the package *tabulizer* by Leeper (2018) is utilized. This package provides bindings for the Tabula PDF extractor, written in Java. In particular, two functions of the library were combined to write a wrapper for the table extraction. First the function *extract_tables()* was used to attempt automized table detection and subsequent extraction. Unfortunately, however, there are quite a few cases in which the automatic table detection failed because some auctions only have one or two bidders. The resulting tables that summarize those auctions have very few rows and thus the automatic detection does not recognize them as tables. Accordingly, if the output of *extract_tables()* is empty, the implemented wrapper calls *extract_areas()*. This function allows the user to specify an area via the R plot-pane to assess where exactly the table is located. Once the location is passed manually, which was necessary for around 10-15% of cases, the extraction works as intended. To finally, obtain the tables the wrapper is used to loop over the PDFs.

2.2 Descriptive Statistics

The data that results from the scraping process is based on an auction level. Meaning that each of the 430 auctions that were scraped between 2015 and 2019 represents one row. The final columns are thus:

- Contract ID
- County
- Letting month
- Letting year
- Contract time
- Number of bidders
- Engineer's estimate
- Award price
- 169 binary variables, representing the bidder identities
- 652 binary variables, representing pair-wise bidder interaction terms
- 258 binary variables, representing the contract description hit words

The interaction terms were generated in order to see, whether, a combination of certain bidders leads to a lower or higher award price. Those terms may also be used to perform unsupervised collusion detection utilizing post-selection inference for ℓ_1-penalized models. This idea is further outlined in Section 5.2. To obtain a concise overview of the data, histograms and bar charts of the variables are provided below.

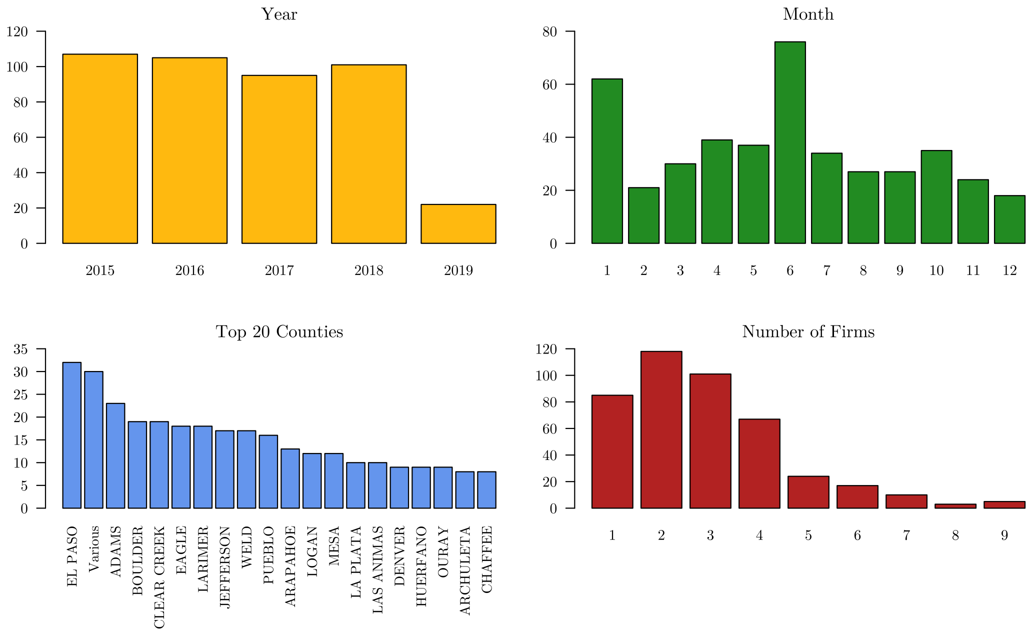

Figure 4: Date, Counties and Number of Firms

We observe, that the sample of auctions in the dataset were held between 2015 and 2019, in which the auction frequency is the highest in January and June. Further we learn, that most contracts are to be completed in El Paso followed by a combination of various counties. In regards to the number of bidders per auction, most of the auctions attract between 1 to 4 bidders, with the maximum being 9 competing firms.

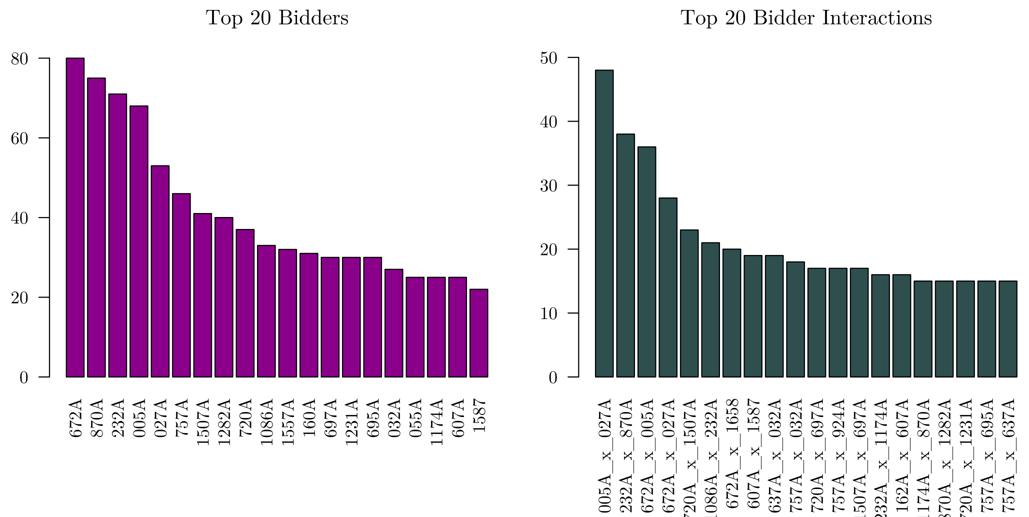

Figure 5: Bidders and their Interactions

The depiction of the top 20 bidders represents the unique identifiers of the firms that submitted the most bids. It is interesting to see that there seem to be a handful of firms that compete in drastically more auctions than others. Additionally, we observe that the two firms that submitted the most bids in the same auctions, compete in slightly more than 10% of the auctions in the dataset.

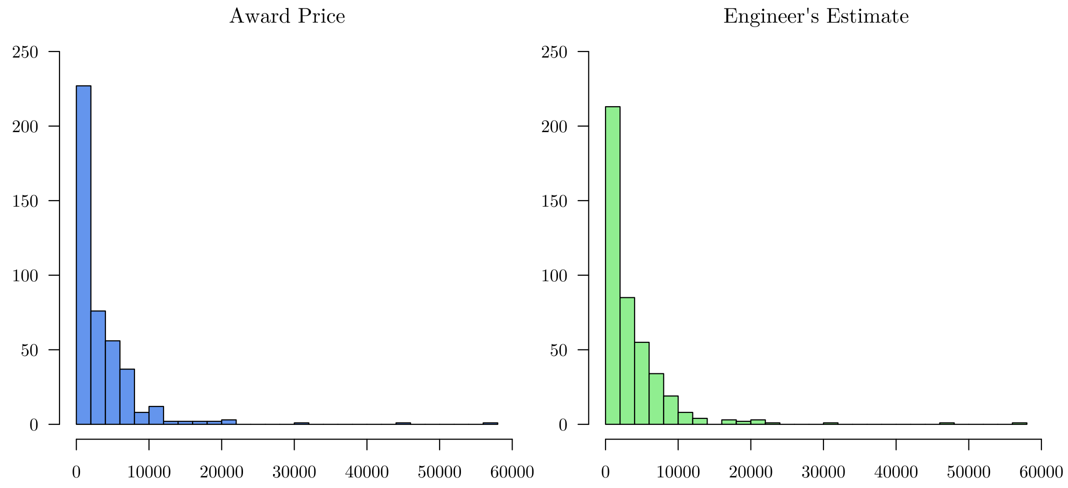

Figure 6: Award Price and Engineer's Estimate

The last two plots, enable us to gain insights about the distribution of the award price and the engineer's estimate. Both variables are right skewed and to the naked eye the engineer's estimate seems to resemble the distribution of the award price quite well. This is a strong indicator for the engineer's estimate as a predictor. The engineer's estimate is also important as it serves as a benchmark for prediction. Every model that can not beat the engineer's estimate in predicting the award price is not particularly useful, this is further discussed in Section 5.

3 Economic Operationalization

Before formalization, a quick overview of the auction design of the Colorado Department of Transportation is presented. This renders modelling assumptions more comprehensible. The CDOT provides a document on their website that summarizes the entire auction process including all preregistration processes, that firms have to traverse before they are eligible to submit bids (CDOT, 2018).
The auction design created by the CDOT is a classical sealed-bid first-price auction, i.e., contracts are advertised for bids online, firms that have been pre-qualified for bidding for a certain contract may submit their bids online or in person. Once a preset date passes a representative of the CDOT opens all submitted bids and announces the lowest bidder, who is awarded the contract. The pre qualification that was mentioned is of vital importance for this thesis. Given that all firms have to be pre-qualified for bidding, one may utilize bidder identity and bidder interactions when predicting award prices. If there was no pre qualification, then the auctioneer would not have knowledge of all potential bidders until no more bids are to be received. At this point there is also no need for a prediction of the award price. Besides a pre-qualification process, the CDOT also implements a reserve price. Every project is assigned an engineer, who is responsible to calculate an engineer's estimate of the cost of completion. Should the lowest bid that was submitted exceed the engineer's estimate by more than 10%, then the lowest bidder is only awarded the contract if a secondary assessment of the CDOT ends with the conclusion that the lowest apparent bid is indeed a reasonable price for execution of the contractual obligations. This secondary process to bypass the initial reserve price was necessary for 17% of the auctions in the dataset that is used in this thesis. In regards to the frequency of bid rejection, for our modelling assumption it is important to know that the secondary assessment process very rarely results in a rejection of the lowest apparent bid (CDOT, 2018).

3.1 First Price Sealed Bid Auctions

For the applied nature of this thesis a formalization of the auction process in not a necessary prerequisite for a good prediction model, however, a sound understanding of all the economic agents at play might yield insights into the potential benefit of a prediction model for participating firms and it may also provide us with expectations for the post-selection inference that is explored in Section 4.1.1. Given that this chapter is not strictly necessary for the reader to understand the main findings of this thesis a basic understanding of game theory is assumed. The interested reader is referred to, Tadelis (2012), who provides sound definitions and explanations of all the required concepts.
The description of bidding in the first-price sealed-bid procurement auction (FPSBPA) is denoted in accordance with Milgrom and Weber (1982) and Schmalensee and Willig (1989). For the following formalization let uppercase letters represent random variables

and their lowercase counterpart a single realisation. Consider an auction environment with n potential bidding firms, all of them submit bids in hopes to procure a contract from the auctioneer. Each of those firms, which are indexed by i observe a real valued signal x_i. One can think of X_i as private information that each firm has about their own cost structure and other private information that a bidder can posses, that influences the value of a contract to a firm. Further, let V represent the characteristics that determine the value of the contract to the firms, that are common knowledge among competing bidders. For illustrative purposes, consider an auction for a contract that involves the repair of a street that leads up a mountain. The information that can be extracted from the contract description is common knowledge, i.e., it influences the value of the contract of every firm. Every bidder is aware that a street, which is steep is more cost intensive to repair. When modelling the bidders behaviour we assume that this is common knowledge, such common values are represented by V. In the auction modelling literature, a distinction is made between an auction environment of private values, i.e., each firm knows their own valuation and is only uncertain about how other firms value the contract. The counterpart to this auction environment is one of pure common values. In this case all bidders have the same common value, however, prior to the execution of the contractual obligations they have no accurate ex ante estimate of the value that may be extracted from completing this contract. This common value paradigm is important when modelling because given that firms have common metrics that make contracts more or less attractive makes the private information that firms may posses relevant for rival bidders and will thus influence the ex post value once all bids are revealed. To continue the formalization of the auction environment, let firm $i's$ payoff be represented by $U_i = u_i(V, X_i, X_{-i})$, where X_{-i} represents the private information possessed by firm $i's$ rivals. Most of the time it is assumed that firms are risk neutral, i.e., expected utility of profit is simply equal to the expectation of a linear function of capital gained, and maximizing it is equivalent to maximizing expected profit itself. The payoff or the value of a contract to a firm may then be defined as the expectation of U_i given the private information of firm i, i.e., $\mathbb{E}(U_i|X_i = x_i, X_{-i} = x_{-i})$. It is important to point out, that in a common values environment the distribution of V conditional on x_i is dependent on the realization of X_{-i}. This means that from the perspective of firm i, other firms may have private information about characteristics of a contract that influence it's value to all competing firms and this also the value in the eyes of firm i. Accordingly, let the bidding strategy be a mapping $\beta : X_i \to \mathbb{R}^+$, that maps the private information of firm i (cost structure, private knowledge about a contract obtained through on site inspections, ...) into the non-negative real numbers. To ensure that this function is invertible, in the literature it is usually assumed that β is a bijection that is strictly increasing in the signals X_i. Again, assuming that X_i represents the cost structure of firm i and other private information, a ceteris paribus decrease in costs is then associated with a decrease in the price of submitted bids. Now that the basic formalities of the auction environment have been set, we may consider the "winner's curse", which assumes the event of winning an auction as an informative event in respect to the value of the contract. The idea is that a firm that manages to procure a contract through an auction must be the lowest bidder, and this also means that no other firm was willing to execute the contractual obligations for the same price. More formally, assume that the realization v is unknown and carries a significant common component. Additionally, we may assume that all firms have varying amounts of private information about the contract. They are aware that other firms have private information but they are uncertain about other firms' valuations of the contract. Further, to simplify the model we assume that the signals x_i are drawn from a common distribution F and that the value of the contract is mainly determined by the common component of bidders' valuations, i.e, w.l.o.g for firm 1, the ex ante valuation of the contract is equal to:

$$v_1(x_1) = \mathbb{E}[V|X_1 = x_1] = \int v f_{V|X_1}(v|x_1)dv.$$

Then in equilibrium, assuming that the bidding function is increasing in the cost realization, the winner is the firm with the lowest cost draw from the distribution but also influenced by the firms private information about the contract. Formally, in the literature this is formalized by stating that the lowest bid is announced by the firm that incurs the lowest signal. This intuitively makes sense as the signal is defined on the real numbers and the lowest signal will thus be identical to the lowest contract execution costs and consequently the highest profit to be made. By Jensen's inequality, given that the maximum is a convex function,

$$\mathbb{E}[\max\{V_i|v\}] \leq \max\{\mathbb{E}[V_i|v]\} = v.$$

This represents the winner's curse as the firm with the lowest signal submits the lowest bid. Further as Milgrom and Weber (1982) state, w.l.o.g, for firm 1 this means:

$$\mathbb{E}[V|X_1 = x, Y_1 > x] < \mathbb{E}[V, X_1 = x].$$

Where Y_1 is the firm with the lowest signal among firm 1's rivals. We thus find that the bidder that incurs the lowest signal associated with the completion of all contractual obligations and therefore the highest potential profit to be extracted from the procurment of the contract should bid less than the firm's ex ante estimate of V_1, simply because the assignment of the contract to the firm as the lowest bidder is an informative event. Especially for firms, that are not as established and might have difficulties to obtain the expertise to properly assess the difficulties associated with completion of a contractual agreement will suffer more severely from the winner's curse. This is relevant for this thesis as an award price prediction that gives a reasonable estimate for the costs associated with the completion of a contract will change the informational asymmetries and thus also lessen the severity of the winner's curse, should the award price prediction be shared with the bidders (Rodríguez et al., 2020).

Now that we have discovered another potential beneficiary of a award price prediction model, we may assess the strategic bidding behaviour of firms. To further simplify the strategic behaviour we may assume that firms signal realisations x_i are drawn from a common distribution, i.e., the value associated with a contract $(V, X_1, ..., X_n)$ are governed by some joint distribution denoted by F. Further, we may assume that the number of firms that submit bids is common knowledge and that the reserve price is not binding. Essentially, this means that the firms believe that bidding above the 110% threshold will not lead to a reissue of the contract but to a contract award in the secondary process previously described. Given that this game is symmetric, we may again focus on firm 1, w.l.o.g. The firm then aims to maximize its expected profit, i.e.,

$$\arg\max_b \ (b - x)\{1 - F[\beta^{-1}(b)]\}^{n-1}.$$

Where $(b - x)$ is the profit that the firm expects if it were to win the contract and the remaining term is the probability that all other competing bidders incur a higher signal realization associated with the execution of the contractual obligation, and thus submit higher bids. Accordingly, by the previously mentioned assumption of the strictly increasing function β, the lowest signal leads to the lowest bid. From the maximization problem, we obtain the following FOC:

$$\begin{aligned} \{1 - F[\beta^{-1}(b)]\}^{n-1} + (b - x)(n - 1)\{1 - F[\beta^{-1}(b)]\}^{n-2}\frac{d\beta^{-1}}{db}\{(-1)f[\beta^{-1}(b)]\} &\stackrel{!}{=} 0 \\ \iff \frac{\{1 - F[\beta^{-1}(b)]\}^{n-1}}{\{1 - F[\beta^{-1}(b)]\}^{n-2}} = (b - x)(n - 1)f[\beta^{-1}(b)]\frac{d\beta^{-1}}{db} \\ \iff 1 - F[\beta^{-1}(b)] = (b - x)(n - 1)f[\beta^{-1}(b)]\frac{d\beta^{-1}}{db} \end{aligned}$$

Knowing that in the bayesian NE, $\beta^{-1}(b) = x$ and $\frac{d\beta^{-1}(b)}{db} = \frac{1}{\beta'(x)}$, we then find:

$$\beta'(x) = \frac{[\beta(x) - x](n - 1)f(x)}{1 - F(x)}$$

This is a differential equation that modern computer algebra systems like mathematica can easily solve to obtain the bayesian NE bid function (Wolfram Research Inc., 2022).

$$\beta(x) = x + \frac{\int_x^\infty [1 - F(t)]^{(n-1)}dt}{[1 - F(x)]^{(n-1)}}$$

We thus learn that the bid resulting from the cost structure of the firm is a surcharge added to the realization of X. Again, we may interpret this as the private information of firm i on the cost of completion of the contractual obligations, e.g., information about the firms cost structure, information about execution costs obtained from on-site inspections, etc. Further, we observe that the surcharge depends on the number of firms in an auction and the joint distribution F. We have to keep in mind, however, that this simple result is based on quite restrictive assumptions. Especially, the risk neutrality of firms is not particularly realistic. Nonetheless, given those simplifying assumptions, the results displayed in section 5 will show, whether, the derived economic intuition of firms behaviour is observable in the data.

4 Methods

This section contains precise descriptions of the methods used to generate the results presented in section 5.

4.1 Elastic Nets

When the number of features in a dataset is significantly larger than the number of observations, elastic net regularization offers a solution to fit linear models using a convex combination of the lasso and ridge penalty. Further the regularization, may be used to better optimize the bias variance trade-off when compared to standard linear regression. The description of the elastic net regularization is denoted in accordance with the seminal paper by Zou and Hastie (2003). Assume that a dataset has n observations and p explanatory variables. As it is the case with the auction dataset, the number of explanatory variables may greatly exceed the number of observations, i.e., $p >> n$. Let $\mathbf{y} = (y_1, ..., y_n)^T$ be the response variable and $\mathbf{X} = [x_1|...|x_p]$ the model matrix, containing the hot encoded explanatory variables. This means that all factor variables have to be converted to individual dummy variables for each level. Subsequently, each column in the model matrix is then centered and standardized. This ensures that the regularization is consistent across variables, since the regularization process is not scale invariant. For any combination of non-negative λ_1 and λ_2, Zou and Hastie (2003), define the elastic net criterion as,

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}) = |\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2 + \lambda_2|\boldsymbol{\beta}|^2 + \lambda_2|\boldsymbol{\beta}|_1,$$

where $|\boldsymbol{\beta}|^2$ represents the ℓ_2 norm of the coefficient vector $\boldsymbol{\beta}$, i.e.,

$$|\boldsymbol{\beta}|^2 = \sum_{i=1}^{p} \boldsymbol{\beta}_i^2$$

and $|\boldsymbol{\beta}|_1$ is equivalent to the ℓ_1 norm, i.e.,

$$|\boldsymbol{\beta}|_1 = \sum_{i=1}^{p} |\boldsymbol{\beta}|.$$

The elastic net estimator $\hat{\boldsymbol{\beta}}$ results from the following minimization problem,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} L(\lambda_1, \lambda_2, \boldsymbol{\beta}).$$

Alternatively, the elastic net penalized regression may also be written as a constrained optimization problem of the least square estimator. Let $\alpha = \frac{\lambda_2}{\lambda_2+\lambda_1}$, then for some fixed

t,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} |\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2,$$

subject to,

$$(1-\alpha)|\boldsymbol{\beta}|_1 + \alpha|\boldsymbol{\beta}|^2 \leq t.$$

The constraint thus encompasses a convex combination of the lasso penalty and the ridge penalty. Accordingly, both ridge and lasso regression are a special cases of an elastic net, where $\alpha = 1$ results in ridge regression and $\alpha = 0$ leaves us with the ℓ_1 penalized lasso regression estimator, initially presented by Tibshirani (1996). In regards to model selection, lasso regression is particularly powerful as it does not shrink candidate variables close to zero, which is the case when utilizing the ridge penalty. In the case of ℓ_1 penalized regression, the obtained $\hat{\boldsymbol{\beta}}$ carries only non-zero components for the variables that were deemed important enough in the cross-validation process used to determine the optimal value of the shrinkage parameter t, assuming $\alpha = 0$. This is troublesome for the frequentist statistician, as it implies that the coefficient vector $\hat{\boldsymbol{\beta}}$ itself is random (Lee et al., 2016). The severity of the implications for frequentist inference is outlined in Benjamini and Yekutieli (2005) and Benjamini et al. (2009).

4.1.1 Post-Selection Inference for ℓ_1-Penalized Models

Having discovered the randomness introduced by shrinkage to exactly zero of some elements in the coefficient vector $\hat{\boldsymbol{\beta}}$, it is important to develop a sound understanding of the source of randomness. The randomness introduced by variable selection is one of choice and not inherent to the parameter value itself. Denoted in accordance with Lee et al. (2016), again we assume p explanatory variables. Further, let $M \subset \{1, ..., p\}$ be the subset of covariates which estimate β_j^M has not been set to zero given the optimal choice of the regularization parameter t. Formally we may write,

$$\{\beta_j^M : M \subset \{1, ..., p\}, j \in M\}.$$

Given $t > 0$, we will only draw inference from the coefficients $\hat{\beta}_j^{\hat{M}}$, i.e., the estimates chosen by the applied ℓ_1 regularization process, $\{\hat{M} : \hat{\beta}_j \neq 0\}$. Thus the randomness is introduced by the choice of $\hat{M}$. Now that the problem inherent with the choice of $\hat{M}$ is established, we may consider a solution to draw valid inference as presented by Lee et al. (2016). Suppose we want to create a confidence interval $C_j^{\hat{M}}$ for a parameter $\hat{\beta}_j^{\hat{M}}$, then fixing α as the desired type I error we want the confidence interval to fulfill:

$$\mathbb{P}(\beta_j^{\hat{M}} \in C_j^{\hat{M}}) \leq 1 - \alpha.$$

Unfortunately, this is obviously undefined for all $j \notin \hat{M}$. Accordingly, to resolve this issue Lee et al. (2016) rely conditional convergence. Since we construct an interval for

$\hat{\beta}_j^{\hat{M}}$ iff $\hat{M}$ is selected, we may condition on this event. Thus we obtain,

$$\mathbb{P}(\beta_j^{\hat{M}} \in C_j^{\hat{M}} | \hat{M} = M) \leq 1 - \alpha.$$

In order to obtain the interval mentioned above, Lee et al. (2016) study the conditional distribution,

$$\boldsymbol{\eta_M^T y} | \{\hat{M} = M\},$$

which allows, in theory, to draw inference from parameters of the more general form $\boldsymbol{\eta_M^T \mu}$. To then characterize the event, $\{\hat{M} = M\}$ for lasso regression the resulting event is a union of polyhedra, i.e., $\{\hat{M} = M, \hat{s}_M = s_M\}$, that conditions on the specified models as well as the signs of the selected covariates. Formally this may be stated as,

$$\boldsymbol{y} \in \mathbb{R}^n : A(M, s_M)\boldsymbol{y} \leq \boldsymbol{b}(M, s_M)$$

and hence when conditioning on the sign as well as the selected variables, it is sufficient to study,

$$\boldsymbol{\eta^T y} | \{A\boldsymbol{y} \leq \boldsymbol{b}\}.$$

Lee et al. (2016), elegantly derive that the resulting conditional distribution is a univariate truncated Gaussian. Subsequently, they use this result to present a test statistic that may be utilized to draw inference about the significance of the parameter estimates $\hat{\beta}_j^{\hat{M}}$. A description of this derivation is beyond the scope of this thesis, the interested reader is referred to Lee et al. (2016) and Taylor and Tibshirani (2016). Where the latter provides an extension to the results of Lee et al. (2016), to generalized ℓ_1 penalized linear models and further also provides the implementation in an R package called *selectiveInference*, by Tibshirani et al. (2019).

4.2 Ensemble Methods

4.2.1 Random Forest

4.2.2 eXtreme Gradient Boosting

4.3 Nested Cross Validation

4.3.1 Logistic PCA

4.3.2 Recursive Feature Elimination

5 Results

5.1 Prediction

5.1.1 Performance Evaluation Metrics

5.2 Unsupervised Colusion Detection

6 Conclusion

References

Benjamini, Y., Heller, R., & Yekutieli, D. (2009). Selective inference in complex research. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, *367*, 4255–71.

Benjamini, Y., & Yekutieli, D. (2005). False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, *100*(469), 71–81.

Benoit, K., Muhr, D., & Watanabe, K. (2021). *Stopwords: Multilingual stopword lists* [R package version 2.3].

CDOT. (2018). *Rules governing construction bidding for cdot public projects.*

CDOT. (2021). *Goods and services procurement presentation for vendors.*

Freitas, W. (2022). *Bizdays: Business days calculations and utilities* [R package version 1.0.11].

Lee, J. D., Sun, D. L., Sun, Y., & Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, *44*(3).

Leeper, T. J. (2018). *Tabulizer: Bindings for tabula pdf table extractor library* [R package version 0.2.1].

Milgrom, P. R., & Weber, R. J. (1982). A theory of auctions and competitive bidding. *Econometrica*, *50*(5), 1089–1122.

Ooms, J. (2020). *Hunspell: High-performance stemmer, tokenizer, and spell checker* [R package version 3.0.1].

Rodríguez, M. J. G., Montequín, V. R., Fernández, F. O., & Balsera, J. M. V. (2020). Bidders recommender for public procurement auctions using machine learning: Data analysis, algorithm, and case study with tenders from spain. *Complexity*, *2020*, 1–20.

Schmalensee, R., & Willig, R. (Eds.). (1989). *Handbook of Industrial Organization* (Vol. 2). Elsevier.

Silge, J., & Robinson, D. (2017). *Text mining with r: A tidy approach* (1st). O'Reilly Media, Inc.

Tadelis, S. (2012). *Game theory: An introduction* (1st ed.). Princeton University Press.

Taylor, J., & Tibshirani, R. (2016). Post-selection inference for l1-penalized likelihood models.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Tibshirani, R., Tibshirani, R., Taylor, J., Loftus, J., Reid, S., & Markovic, J. (2019). *Selectiveinference: Tools for post-selection inference* [R package version 1.2.5].

Wickham, H. (2022). *Rvest: Easily harvest (scrape) web pages.*

Wolfram Research Inc. (2022). *Mathematica, Version 13.1* [Champaign, IL, 2022].

Zou, H., & Hastie, T. (2003). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.