Predicting Award Prices of First Price Sealed Bid Procurement Auctions

Fabian Blasch

Supervisor: Dr. Katharina Fenz

July 20, 2022

Contents

1 Introduction

The importance of public procurement becomes quickly apparent when looking at the shere volume of contracts that is awarded through public procurement auctions. The authorities of the European Union for example spent around 14% of their GDP on public procurement 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 prcorurment of street and bridge building and repair 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).

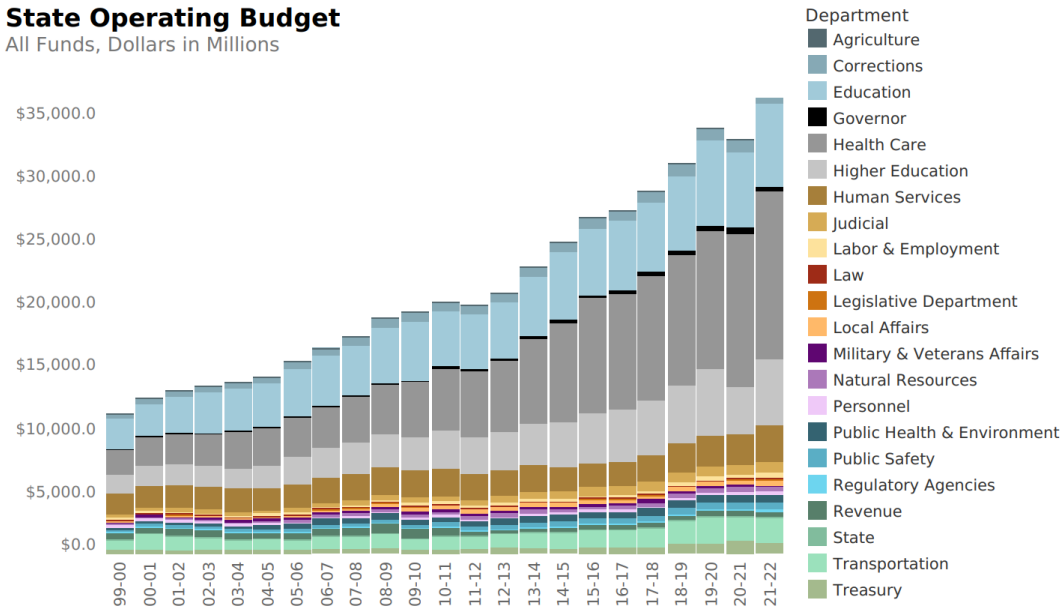


Figure 1: State of Colorado Budget

Gieven that the contracts awarded through procurement auctions are such a substantial part of the budget, every potential improvement to the process in place may greatly increase the efficiency of the tax payers money spent. Accordingly, a closer examination of the process including a prediction model for the auctions' award prices may support public procurment 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 propcurement 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

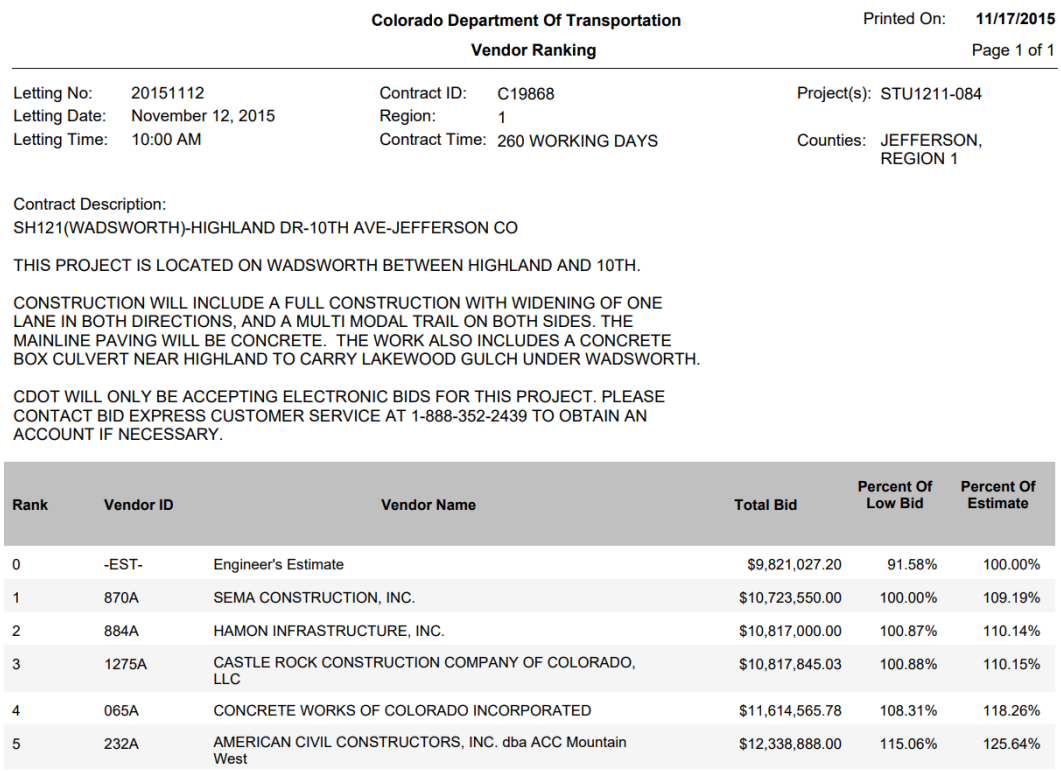
are published on the official website of the CDOT. In particular, four different model types with different preprocessing schedules are compared in terms of their predictive power. Standard linear regression, lasso regression, random forests and an eXtreme gradient boosting model. To assess, whether, the combination of certain bidders leads to higher award prices, recently discovered post-selection inference methods are utilized. The remaining paper is strucutured 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 described, once from the perspective of the auctioneer and also from the perspective of the competing firms. The following chapter then covers the methods used, not only in respect to the different models used for prediction but also for the different pre-procerssing shedules, that are applied and the post-selection inference that is used. The thesis then 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 documents 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	Contract ID:	C19868	Project(s):	STU1211-084
Letting Date:	November 12, 2015	Region:	1		
Letting Time:	10:00 AM	Contract Time:	260 WORKING DAYS	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 fullfilled. 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 substracted from the difference betwen 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 will only take 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, stopwords were removed from the list of tokens. Stopwords 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 stopwords, depending on the context different libraries of stopwords 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 stopwords 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 stopwords, 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 determin 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 stopwords 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.

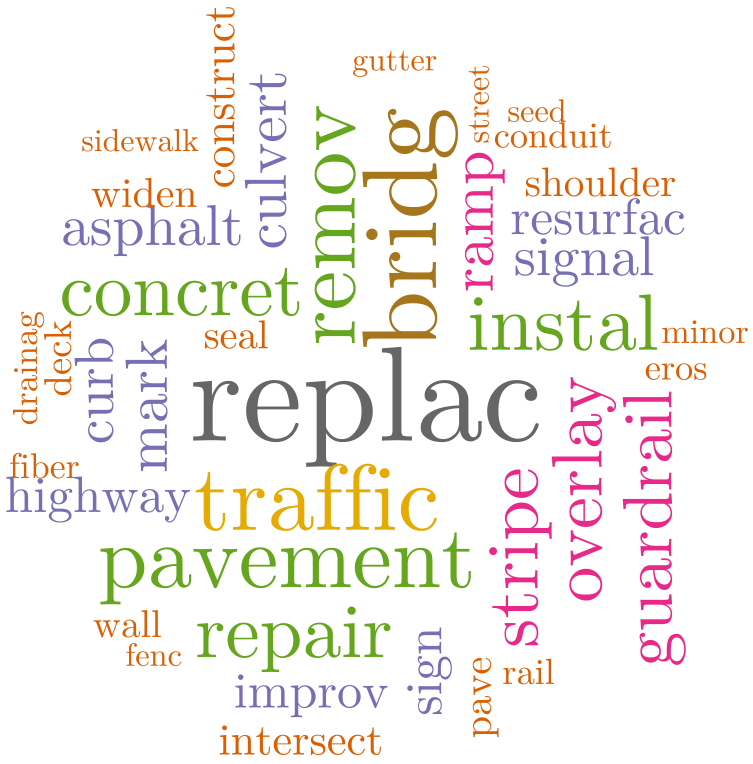


Figure 3: Top 40 Stemmed Description Words

2.1.2 Tabular Data

As displayed in Figure 2, the table in each of the auctions' PDFs contains information on the submited 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 specify 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 sraping 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 hitwords

The interactionterms 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 barcharts of the variables are provided below.

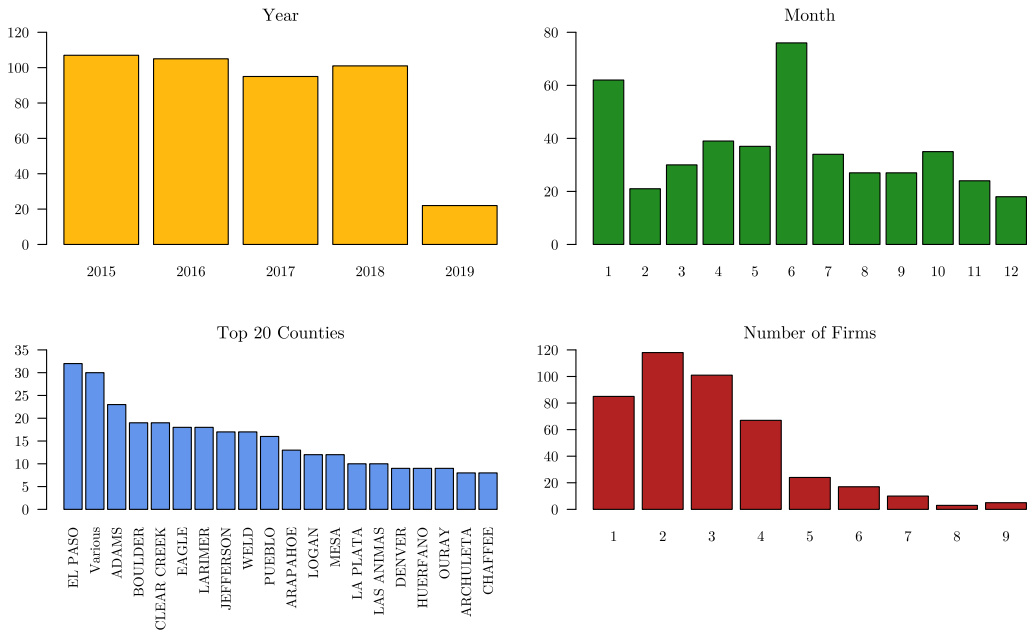


Figure 4: Date, Counties and Number of Firms

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 multiple counties. In regards to the number of bidders per auction, most of the auctions have 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 drastrically 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.

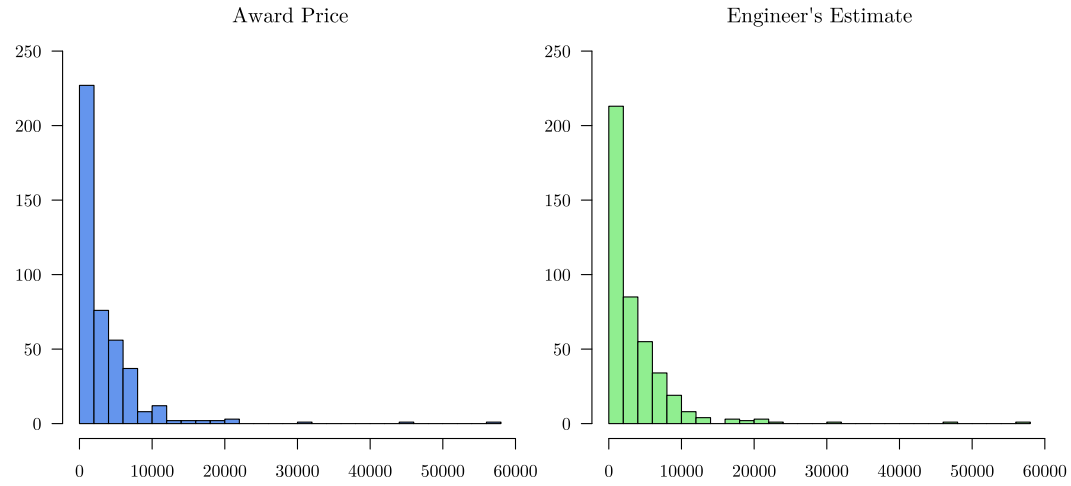


Figure 6: Award Price and Engineer's Estimate

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

3.1 Auctioneer

3.2 Firms

4 Methods

4.1 Elastic Nets

4.1.1 Post-Selection Inference for ℓ_1-Penalized Models

4.2 Ensemble Methods

4.2.1 Random Forests

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

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

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

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

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

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.

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

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