Research Project

2022-2023

Bidders Recommender for Public Procurement Auctions

Using Machine Learning

IASMINA-OANA SILASCHI
RESEARCH PROJECT

Table of Contents

1. Classification

1.1. ACM:

J.1

1.2. MSC:

68T09

91-11

1.3. Key words:

machine learning

artificial intelligence

ranking

economic data

data processing

public procurement auctions

2. Abstract

This application will rank bidders considering the best fit for a public auction. Taking into consideration the persistent issues regarding Romania's integrity in the case of contracting companies for public procurement auctions, this application could validate if indeed the best decision has been made in auctions from the past, as well as be helpful regarding this matter in future tenders.

Even though this subject has already been researched, it has never been specifically tailored for tenders from Romania and using a dataset of companies from this country. Therefore, considering the specific legislation of this European country, it will have original results. Moreover, besides computing the profile of the ideal candidate for the contracting authority, my algorithm will also rank the bidding companies according to their compatibility.

AI can be used in various ways to assist the public procurement auctions process, such as analyzing large amounts of data to make predictions about bid success, automating the evaluation of bids, identifying patterns in previous auction results, and detecting fraud or misconduct. These functionalities can increase the efficiency, fairness, and transparency of the process, and help suppliers to be more competitive.

A real case study is proposed where a dataset of public procurement tenders in Romania is collected, including information about the contracting authorities, companies, and bids. The sample consists of 47,974 contracting authorities, 289,472 companies and 42,474 bids. With this dataset, experiments can be conducted to evaluate the performance of a machine learning-based ranking tool for the tenders, which can help to improve the efficiency, fairness and transparency of the procurement auction process and provide insights for future tenders.

3. Introduction

3.1. Public Procurement Auctions Bidding

Public procurement auctions are a common way for governments and other public entities to purchase goods and services. They involve the submission of bids by suppliers, and the selection of a winning bid based on a set of predetermined criteria.

Artificial intelligence (AI) can be used to assist with the bidding process in several ways. For example, AI algorithms can be used to analyze large amounts of data and make predictions about which bids are likely to be successful. This can help suppliers to optimize their bids and increase their chances of winning.

AI can also be used to automate the process of evaluating bids, by analyzing the various criteria (such as price, quality, and delivery time) and determining which bid offers the best value for money. This can save time and resources for the organization running the procurement auction.

Another way AI can assist the process is by identifying patterns in the data of previous auction results, allowing bidders to adjust their bid's strategy to increase the chances of winning.

In addition to these functionalities, AI-based tools can also be used to detect fraud or misconduct during the procurement auction process, by analyzing patterns of behavior and detecting any anomalies or suspicious activity.

Overall, using AI in public procurement auctions can increase efficiency, fairness, and transparency of the process while also helping suppliers to be more competitive.

3.2. Ranking using Machine Learning

Ranking is the process of ordering a set of items according to some criteria. Machine learning (ML) can be used to create models that can predict the ranking of items based on input data.

Ranking bidders for public procurement auctions using machine learning (ML) involves training a model to predict the likelihood of a bid being successful, based on a set of input features such as the bid amount, the qualifications of the bidder, and the specific terms of the bid.

To train such a model, a dataset of past procurement auctions with information about the bids and the outcome of the auction is needed. This dataset would include features about the bidders such as their past performance, financial stability, and relevant experience in the area of the goods or services being procured.

Based on the data, various machine learning models can be built and trained. Commonly used algorithms for this kind of problem include linear regression, decision trees, and random forests. These models can then be used to predict the likelihood of success for new bids, based on the input features provided.

Additionally, the ranking can be further improved by the use of more advanced algorithm like Gradient Boosting Decision Trees, Random Forest and Support Vector Machine, or even deep learning based neural networks.

The goal of the model is to accurately predict the outcome of the auction, and to optimize the model's ability to make accurate predictions. This can be measured using metrics such as accuracy, precision, recall, and mean average precision. These predictions can be used to help suppliers optimize their bids and increase their chances of winning.

Moreover, the use of machine learning in ranking bidders can also improve the transparency and fairness of the procurement auction process. Since the ML model’s decision is based on data and the specific rules defined and can provide explanations of the decision and can be audited for any bias.

4. Related Work

Article 1: (Casalegno, 2022)

Abstract:

In this post, by "ranking" we mean sorting documents by relevance to find contents of interest with respect to a query. Ranking models typically work by predicting a relevance score for each input. Once we have the relevance of each document, we can sort (i.e., rank) the documents according to those scores. The scoring model can be implemented using various approaches.

Article 2: (Chen, 2009)

Abstract:

Learning to rank has become an important research topic in machine learning. While most learning-to-rank methods learn the ranking functions by minimizing loss functions, it is the ranking measures that are used to evaluate the performance of the learned ranking functions. In this work, we reveal the relationship between ranking measures and loss functions in learning- to-rank methods, such as Ranking SVM, RankBoost, RankNet, and ListMLE. We show that the loss functions of these methods are upper bounds of the measure- based ranking errors. As a result, the minimization of these loss functions will lead to the maximization of the ranking measures.

Article 3: (Liu)

Abstract:

Learning to rank for Information Retrieval (IR) is a task to automatically construct a ranking model using training data, such that the model can sort new objects according to their degrees of relevance, preference, or importance. Many IR problems are by nature ranking problems, and many IR technologies can be potentially enhanced by using learning-to-rank techniques. The objective of this tutorial is to give an introduction to this research direction. Specifically, the existing learning-to-rank algorithms are reviewed and categorized into three approaches: the pointwise, pairwise, and listwise approaches.

Article 4: (Machine learning (ML) applications: ranking, 2022)

Abstract:

Ranking is a type of machine learning that sorts data in a relevant order. Companies use ranking to optimize search and recommendations.

Article 5: (S. Pandey, 2019)

Abstract:

Information retrieval is the topic area in which much research has been done and is still going on. The rapidly growing web pages make it very crucial to search up to date documents. In continuation of research works on learning to rank, this research focuses on implication of machine learning techniques for IR ranking.

5. Literature Review

Article 1: (Casalegno, 2022)

Relevance:

This article is relevant because it focuses on how to implement Machine Learning models for Learning to Rank, explaining how this pertains in many applications. Also, the scoring model is thoroughly explained in an intuitive way.

Article 2: (Chen, 2009)

Relevance:

I believe this article is relevant to my research project because it includes various mathematical functions and computations which are vital to the implementation of this subject.

Article 3: (Liu)

Relevance:

This is a very comprehensive article which amasses plenty of useful insights for this subject, and it will serve as a great source to find information.

Article 4: (Machine learning (ML) applications: ranking, 2022)

Relevance:

This article includes suggestive drawings to help better understand the topic, and it is short and concise, so I believe it is a great starting point to dive into the researched subject.

Article 5: (S. Pandey, 2019)

Relevance:

The techniques that can be used are explained very well, as well as compared to one another, thus being a very interesting and helpful article.

6. Material and Methods

6.1. Dataset

Table 1: **Bids**

noticeNo
contractValue
title
currencyCode
publicationDate
countyCode
contractingAuthorityName
cpvCodeName
cpvCode
cpvCodeType
cNoticeNo
cNoticeEstimatedContractValue
cNoticePublicationDate
cNoticeTitle
organizationName
organizationId

Table 2: **Contracting Authorities**

contractingAuthorityId
name
CUI

Table 3: **Oganizations**

organizationId
name
CU

Data description

a) Noticeno - the number of the award announcement related to the purchase in SEAP (helps to identify easier procedures for various analysis flows)

b) contractvalue - the contractual value introduced by the contracting authority in the award notice

c) Titles - the title given by the contracting authority of the procedure in the award announcement (text description of the main object of the procedure)

d) Currencycode - procedure currency (RON, EUR, USD)

e) Publicationdate - Date of publication publication

f) Countycode - the county of the contracting authority

g) Contractingauthorityname - the name of the contracting authority, as it appears in SEAP

h) Cpvcodename - the name of the COMMON CODE PROCUREMENT VOCABULUY (CPV) used in the procedure (e.g. afforestry services)

i) CPVCODE-Numeric Code of CPV (cf. European nomenclature, e.g. 77231600-4)

j) CpvCodetype - only works procedures (code 2) and services (code 3), not product supply, were selected, the supply is less relevant to the context of the project

k) Cnototiceno - the number of the announcement of participation related to the award announcement above (helps to identify easier procedures)

l) cnototestematedcontractvalue - the estimated value of the contract in the participation announcement (the award results resulting from the competition between the bidders)

m) cnototicepublicationdate - the date of publication of the participation announcement related to the award announce

n) Cnotototitle - the title given by the contracting authority of the procedure in the participation announcement (may be more object / descriptive than that put in the award announcement)

o) Organizationname - the winner or winners of the procedure (can be consorti)

p) Organization - the Code or the internal numeric codes of Spend.ro for winners (the same winner can have several codes)

6.2. Method

The proposed solution consists of three main phases: forecasting, aggregation, and searching.

1. Forecasting phase:

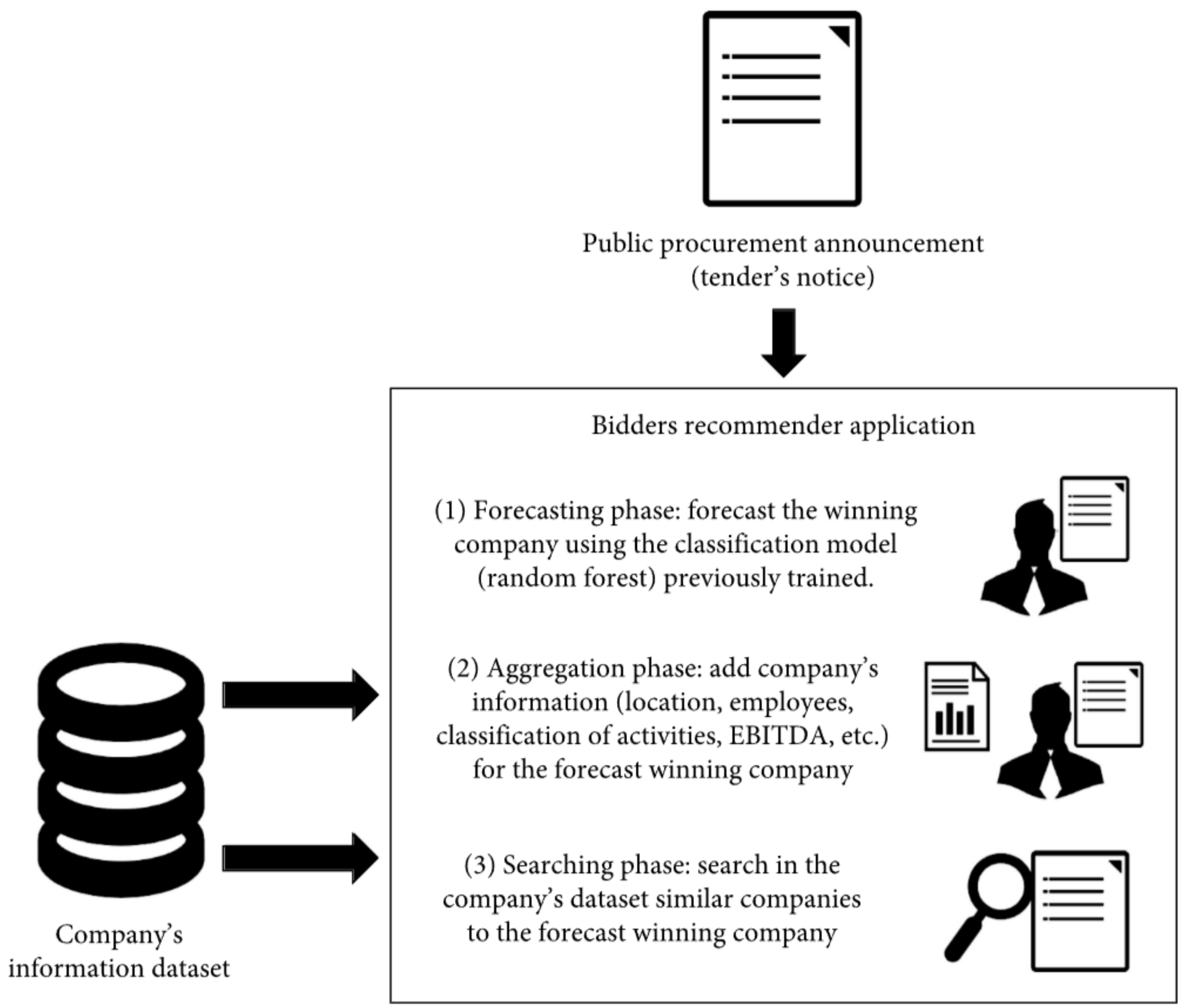
   Forecast the winning company using a previously trained AI.

2. Aggregation phase:

   Add the company's information for the forecast winning company.

3. Searching phase:

   Search in the dataset similar companies to the forecasted winning company and rank them according to their compatibility.



6.3. Algorithm

Random Forest (RF) is an ensemble learning method that can be used for classification or regression problems. It was introduced by Breiman in 2001 and is based on the idea of creating multiple decision trees at training time and then combining their predictions to make a final prediction.

One of the key advantages of RF is that it is a versatile algorithm that can be applied to a wide range of problems, including large-scale problems. It is also able to handle different types of data, such as continuous, categorical, and ordinal data. Additionally, RF can handle missing values, high-dimensional data and correlated features.

Another key advantage of RF is that it can help to reduce the risk of overfitting, which occurs when a model is trained too well on the training data and performs poorly on new, unseen data. This is achieved by creating multiple decision trees, each of which makes a prediction based on a random subset of the features. By averaging the predictions of the individual decision trees, RF is able to reduce the variance and improve the generalization performance.

In recent years, RF has been applied with remarkable success in tender datasets. RF can be used to analyze the data of previous auctions, and identify patterns or relationships that can be used to improve the chances of winning in future auctions. It can be used to predict the outcome of a particular auction, and to identify the factors that are most important for a successful bid. Furthermore, RF can be used to identify fraud or misconduct, by analyzing patterns of behavior and detecting any anomalies or suspicious activity.

Overall, RF is a powerful and versatile algorithm that offers a range of advantages when applied to public procurement datasets. It can be useful to identify patterns in the data, predict the outcome of an auction and identify the factors that are most important for a successful bid, and detect fraud or misconduct.

Simplified algorithm of RF for classification summarized:

(1) For $b = 1$ to B (number of trees):

(a) Draw a bootstrap sample $\mathbb{Z}^*$ of size N from the training data.

(b) Grow a random forest tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size $n_{\min}$ is reached.

(i) Select m variables at random from the p variables.

(ii) Pick the best variable/split point among the m.

(iii) Split the node into two daughter nodes.

(2) Output the ensemble of trees $\{T_b\}_1^B$.

To make a prediction at a new point x, let $\widehat{C}_b(x)$ be the class prediction of the $b-$th random forest tree. Then, $\widehat{C}_{rf}^B(x) = \text{majority vote}\,\{\widehat{C}_b(x)\}_1^B$.

7. Experiments

A real case study:
I will take a tenders list from Romania, a list of the organizations as well as a list of contracting authorities in order to perform experiments.

The sample consists of 47,974 contracting authorities, 289,472 companies and 42,474 bids.

With this dataset, there are several potential experiments that could be conducted to evaluate the performance of the machine learning-based ranking tool described in the abstract. Some examples of experiments that could be conducted include:

1. Building and training a machine learning model: Using the dataset, there could built and trained a machine learning model that is able to predict the likelihood of a bid being successful, based on input features such as the bid amount, the qualifications of the bidder, and the specific terms of the bid.
2. Model Evaluation: After the model is trained, you could evaluate its performance by testing it on a hold-out dataset of bids and comparing its predictions to the actual outcomes of the auctions. This could be done using metrics such as accuracy, precision, recall, and mean average precision.
3. Performance comparison: Once you have a robust model, you could compare the performance of your algorithm with traditional methods such as linear regression, decision trees, or random forests, or other ML models like Gradient Boosting Decision Trees, Random Forest and Support Vector Machine, to see how well it performs.
4. Model's robustness: After you compare your model with traditional ones, you could further evaluate its robustness by testing it on different subsets of the data, such as bids from different regions or bids for different types of goods and services. This would help to determine how well the model generalizes to new cases.
5. Explanation of the decision: Finally, you could use techniques such as LIME, SHAP or other methods for model interpretability to understand the decision of the model, that is why it gave certain ranking to a company and how that company's feature contribute to the decision of the model.

These are some examples of experiments that could be conducted, but it's important to note that the specific details of the experiments will depend on the goal and constraints of the project and the characteristics of the dataset.

7.1. Data Processing

Data preprocessing is an important step in machine learning, as it ensures that the data is in a format that can be easily used for training and evaluating models. In the case of a dataset of public procurement tenders, data preprocessing may be necessary due to human errors such as incorrect formatting, wrong values, or empty fields.

Data preprocessing of the tender dataset is necessary due to the fact that information has not been verified automatically to correct human errors, such as incorrect formatting, wrong values, empty fields, and so on. Data preprocessing can be divided into the following 5 consecutive tasks: extraction, reduction, cleaning, transformation, and filtering.

1. Extraction: This step involves extracting the relevant information from the dataset. This might include identifying and extracting the specific columns or fields that contain the information that is needed for the analysis.
2. Reduction: This step involves reducing the size of the dataset by removing duplicates or irrelevant data. In this case, after the extraction step, the sample reduced from 289,472 companies to 230,770 tenders, it is likely that some companies have multiple tenders or entries in the dataset.
3. Cleaning: This step involves cleaning the data by fixing errors and inconsistencies. This might include correcting errors in formatting, replacing missing values, or removing outliers. In this case, a specific algorithm to check the integrity of the VAT numbers of the companies was used to filter them and ensure the accuracy of the data.
4. Transformation: This step involves transforming the data into a format that can be easily used for analysis. This might include converting data into a different format, such as from string to numerical values or normalizing the data.
5. Filtering: this is an optional step that can be applied to further remove unnecessary data or according to some specific rules, in this case, the use of the algorithm to check the integrity of the VAT numbers can be considered as a filtering step.

After data preprocessing, the number of companies was reduced from 289,472 to 230,770 tenders.

7.2. Statistical Analysis

Statistical analysis is a powerful tool that can be used to understand and describe the characteristics of a dataset. In the case of a dataset of public procurement tenders, statistical analysis can be used to describe the data quantitatively and understand the relationships between the different variables.

Firstly, the most relevant information of the tender dataset will be explained, quantitatively. Secondly, the company dataset will also be explained, as well as the contracting authorities dataset, and, finally, the correlations between all three datasets will be analyzed.

There are several different types of statistical analysis that can be performed on the tender dataset, depending on the specific research question and the characteristics of the data. Some examples of statistical analysis that can be performed include:

1. Descriptive statistics: This type of analysis is used to summarize the main characteristics of the dataset. This might include calculating measures such as the total number of tenders, the mean, median, and maximum values, and the standard deviation.
2. Exploratory data analysis (EDA): This type of analysis is used to explore the dataset and identify patterns or trends. This might include creating visualizations such as histograms, box plots, and scatter plots.
3. Inferential statistics: This type of analysis is used to make inferences about a larger population based on a sample of the data. This might include calculating confidence intervals or testing hypotheses.
4. Hypothesis testing: Specific hypothesis testing can be performed depending on the research question, such as if the distribution of bids is normal, or if there is a relationship between the price and the success of the bid.
5. Correlation analysis: This type of analysis can be used to identify the relationships between different variables in the dataset. This might include calculating correlation coefficients such as Pearson, Kendall, or Spearman correlation.
6. Clustering or segmentation: This type of analysis can be used to group similar observations in the dataset together, can be applied to identify different bidding strategies or company's sectors.

It is worth noting that these statistical analysis are not mutually exclusive, and multiple techniques can be combined to get a comprehensive view of the data. Furthermore, it is important to keep in mind the assumptions and limitations of each method, as well as their interpretability.

8. Conclusion

The conclusion of this research is that the application described is a machine learning-based tool that is designed to rank bidders in public procurement auctions in Romania.

It aims to address the issue of integrity in Romania's public procurement auctions by validating the decisions made in past auctions and providing guidance for future tenders.

The application is based on a dataset of Romanian companies, and it takes into account the specific legislation of the country.

Furthermore, it is supposed to be tailored to Romania specific tenders and is expected to produce original results.

Additionally, the algorithm will not only compute the ideal candidate for the contracting authority but also rank the bidding companies according to their compatibility.

9. Bibliography

1. A rank-and-compare algorithm to detect abnormally low bids in procurement auctions. (2012). In L. D. Pier Luigi Conti, *Electronic Commerce Research and Applications* (pp. 192-203).

2. Application of artificial intelligence in control systems of economic activity. (2019). In O. Melnychenko.

3. Casalegno, F. (2022). Learning to Rank: A Complete Guide to Ranking using Machine Learning. *Towards Data Science*.

4. Chen, W. &.-Y. (2009). Ranking Measures and Loss Functions in Learning to Rank. *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009*, (pp. 315-323). Vancouver, British Columbia, Canada.

5. Dragan Pamučar, D. B. (2022). Application of neuro-fuzzy system for predicting the success of a company in public procurement. In *Decision making: Applications in Management and Engineering* (pp. 135-153).

6. Laura Abrardi, C. C. (2021, July 26). *Artificial intelligence, firms and consumer behavior: A survey.* Retrieved from Wiley Online Library: https://onlinelibrary.wiley.com/doi/full/10.1111/joes.12455

7. Leon Yang Chu, H. N. (2020). Position Ranking and Auctions for Online Marketplaces.

8. Lisa Chever, S. S.-B. (2016). The law of small numbers: investigating the benefits of restricted auctions for public procurement.

9. Liu, T.-Y. (n.d.). *Learning to Rank for Information Retrieval.* Retrieved from http://didawiki.di.unipi.it/lib/exe/fetch.php/magistraleinformatica/ir/ir13/1_-_learning_to_rank.pdf

10. *Machine learning (ML) applications: ranking.* (2022, Mar 17). Retrieved from DEV: https://dev.to/mage_ai/machine-learning-ml-applications-ranking-238d

11. Manuel J. García Rodríguez, V. R. (2020, November 25). *Bidders Recommender for Public Procurement Auctions Using Machine Learning: Data Analysis, Algorithm, and Case Study with Tenders from Spain.* Retrieved from Hindawi: https://www.hindawi.com/journals/complexity/2020/8858258/

12. Manuel J.García Rodrígueza, V. R.-M.-P. (2022). Collusion detection in public procurement auctions with machine learning algorithms. In *Automation in Construction.*

13. Milutin ŽIVKOVIĆ, ,. M. (n.d.). MULTICRITERIA APPROACH TO PUBLIC PROCUREMENT FOR. In *Fiability & Durability* (pp. 181-187).

14. Montequín, V. R. (2019). Public Procurement Announcements in Spain: Regulations, Data Analysis, and Award Price Estimator Using Machine Learning.

15. S. Pandey, I. M. (2019). Information Retrieval Ranking Using Machine Learning Technique. *Amity International Conference on Artificial Intelligence (AICAI)*, (pp. 86-92).