NOVA IMS Information Management School

MDSAA

Master Program in
Data Science and Advanced Analytics

Decoding Success with Zero-Inflated and Hurdle Models: Unveiling the Winning Strategies in Portuguese Public Procurement Activity

Evidence from Portugal

Leonor Matias Porto

Dissertation

presented as partial requirement for obtaining the Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

DECODING SUCCESS WITH ZERO-INFLATED AND HURDLE MODELS: UNVEILING THE WINNING STRATEGIES IN PORTUGUESE PUBLIC PROCUREMENT ACTIVITY

by

Leonor Matias Porto

Dissertation presented as partial requirement for obtaining the Master's Degree in Advanced Analytics with major in Business Analytics

Advisors: Bruno Miguel Pinto Damásio and Flávio Luís Portas Pinheiro

July 2023

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Leonor Matias Porto

Lisboa, 15th July 2023

ACKNOWLEDGEMENTS

The completion of this dissertation marks the end of several months of dedicated effort, undertaken while balancing a full-time work commitment. For this reason, I would like to express my gratitude to the people who have supported me throughout this journey.

First and foremost, I am thankful to my parents for all the support they have given me, the opportunities they have provided, and their role in contributing to the person I am today. They have been crucial to my success and in the achievements of my milestones.

I am also extremely grateful to my boyfriend, Tomás, for his constant support, encouragement and help throughout this work, especially during the most challenging moments. His presence and understanding have been essential to my success.

I extend my appreciation to my advisors, Prof. Bruno Damásio and Prof. Flávio Pinheiro, for never giving up supporting and guiding me. As well as for all the knowledge and motivation transmitted, which has also been central to the completion of this master thesis.

Lastly, I would like to express my gratitude to my friends and work colleagues for their continuous encouragement, and tips, which have also played a significant role in inspiring me to reach the finish line of this academic journey.

ABSTRACT

Public procurement plays a vital role in promoting economic development, employment, innovation, and sustainability within the private and public sector. Traditionally, public procurement has been primarily based on lowest-price bids, but recent shifts have emphasized the need for additional evaluation criteria, such as quality, resources capacity, financial stability, experience in the market, innovation, sustainability, and contribution to the society. Therefore, this study aims to predict the success of Portuguese companies in public tender procedures and identify the key factors influencing their success. Additionally, it examines the factors of influence within each contract type to uncover potential variations across different types of contracts. The study makes use of public procurement data from the BASE portal in Portugal, along with business information of the corresponding participants in public procurement, obtained from the global database ORBIS. The combination of these two datasets forms the comprehensive dataset used for analysis in this study. To measure companies' success in these procedures, the number of public tenders won by each company per year is predicted using count data regression methodologies, considering the discrete nature of the response variable. Advanced models, such as Zero-Inflated and Hurdle models, are employed to effectively handle excess zero values and improve prediction accuracy. The model's evaluation indicates that these models outperform traditional models in addressing the overdispersion and high variance observed in the data. The results allow to identify and quantify the key factors that significantly influence the success of companies in the public tender procedure within each contract type. Overall, it shows that companies' size and experience in the public tender's activity are one of the key factors in the success of winning public tenders. Moreover, the results also shown that the relevance and impact of the different factors studied, which also includes, resources capacity, market experience, profitability and financial stability can vary across contract types. This comprehensive understanding of the determinants of success in public tenders provides valuable insights for companies, enabling them to tailor their strategies and improve their competitiveness in the market. These insights also provide benefits for public authorities by helping to elaborate more effectively the public procurement award criteria and develop targeted policies that support the growth and sustainability of companies facing challenges in the market. These efforts foster a competitive market environment that encourages innovation, economic development, and fair distribution of opportunities.

KEYWORDS

Public Procurement; Count Data Regression; Zero-Inflated models; Hurdle models

Sustainable Development Goals (SDG):

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AIC	Akaike Information Criterion
CPV	Common Procurement Vocabulary
EBIT	Earnings Before Interest and Taxes
EBITDA	Earnings Before Interest, Taxes, Depreciation, and Amortization
EU	European Union
GDP	Gross Domestic Product
HNB	Hurdle Negative Poisson
HP	Hurdle Poisson
IMPIC	Instituto dos Mercados Publicos, do Imobiliario e da Construcao
IQR	Interquartile Range
LL	Log-Likelihood
LRT	Log-Likelihood Ratio Test
MAD	Median Absolute Deviation
MAE	Mean Absolute Error
MEAT	Most Economically Advantageous Tender
MLE	Maximum Likelihood Estimation
NB	Negative Binomial
P	Poisson
PL	Profit & Loss
PMF	Probability Mass Function
PPC	Public Procurement Code
RF	Random Forest
ROA	Return On Assets
ROE	Return On Equity
ROCE	Return On Capital Employed
RMSE	Root Mean Square Error

SME Small and Medium Enterprise

SMV Support Vector Machine

ZI Zero Inflated

ZINB Zero Inflated Negative Binomial

ZIP Zero Inflated Poisson

1. INTRODUCTION

1.1. PUBLIC PROCUREMENT

Public Procurement is defined as the purchase of goods and services by a public institution from an external source (OCDE, 2013). Therefore, Public Procurements are a way of the government to apply its public expense, and it has an important role in the economy of a country once it allows to satisfy needs in the material and services assistance of public institutions and distribute money through the economy (Martins & Damásio, 2020). Thus, public procurement policy goal is to foster regional and companies' development, employment, innovation, sustainability, and efficient use of economic resources (Sarter et al., 2014; Uyarra & Flanagan, 2010).

Given its relevance for the economy in the European Union (EU) countries, the award of public procurement process is required to comply with the principles of non-discrimination, equal treatment, free competition, mutual recognition, proportionality, and transparency (Dotoli et al., 2020; Grzyl et al., 2018) with the aim to reduce fraud and corruption and legal administrative barriers to participation in public procurement (European Commission, 2018).

To make sure these principles are followed, there is legislation that sets detailed specifications regarding on how public procurement should be processed in the EU countries (Glas & Eßig, 2018). Therefore, the national rules of these countries must respect the general principles of EU laws for all the tenders that are published in the Official Journal of the European Union, and whose values must exceed a specific threshold. For the tenders that do not achieve these thresholds, national rules apply (Telles, 2013).

In case of Portugal, public procurement is celebrated under the Public Procurement Code (IMPIC, 2008), and as stated in 1st-A article, during the contract's implementation, the general principles deriving from the Portuguese Constitution, the European Union's treaties, and the Code of Administrative Procedure most be under compliance. According to the 16th article of the PPC there are seven different types of contracting procedures that can be categorized into two groups: non-competitive procedures and competitive procedures. In non-competitive procedures the contracting authority is entitled to directly invite a one or more pre-selected entities to submit a tender, while in a competitive procedure the contracting entity announces its intentions on both *Diário da República* and the Official Journal of the European Union so that possible contractors can apply to the proposal by submitting their bids. In Portugal, the competitive procedure has gained more relevance in most recent years, with a weight of 61,2% in the total contract amount in the year of 2021 (Técnica, 2021).

In terms of awarding criteria, the 74th article states that it's allowed to use only two criteria for contract award purposes: price criteria, where the price evaluation is the only factor to consider in the award decision; or best quality-price criteria, also known as the Most Economically Advantageous Tender (MEAT), which considers a combination of factors and subfactors as criteria for the award decision.

1.2. MOTIVATION

Governments procure a large amount of goods and services to help them implement policies and deliver public services. Thus, the public procurement expenditure as a percentage of the GDP in EU countries, over the last decade, increased from 13,1% of GPD in 2008 to 14,9% of GDP in 2019 (OECD, 2021). In case of Portugal, government spends between 9% to 11% of the GDP (OECD, 2021). This reflects the growing awareness of public procurement as an innovative policy tool and the high interest from both policy makers and researchers (Obwegeser & Müller, 2018).

Nowadays, to determine its survival and development, companies need to continuously keep up with the market, by facing challenges of continuous maintenance and improvement of the level of the services and/or goods provided, technological innovation and competition evaluation (Rietveld & Schilling, 2021). Therefore, is important that companies participate in a competitive market and hence face the necessity of continuing assessment. In particular, these challenges can be even higher for the Small and Medium Enterprises (SMEs), which represent a crucial role in most economies, given their lower capacity and resources, which creates difficulty to compete on equal terms with the bigger companies (Björkegren, 2022). For this reason, public procurement represents an important role in the private sector, given that it's characterized by a number of policy goals that promote innovation, conservation of the environment, support of national enterprises and the increase of SMEs' involvement, while managing efficiently the use of public resources (Baranovsky et al., 2020; Glas & Eßig, 2018; Grzyl et al., 2018; Lenderink et al., 2019).

Traditionally, public procurement awarding process has been based on the lowest-price bid (Ballesteros-Pérez et al., 2015), however in most recent years the awareness of non-price criteria to rank bidders' proposals increased in EU countries, considering several aspects such as natural, human, social and economic factors. Therefore, relevance of the price in public procurement awarding has been decreasing, given that focusing on price criteria to reduce costs might have negative consequences to meet public and state needs, since this might lead to quality issues, delays, additional costs and an imperfect competition in the market (Baranovsky et al., 2020; Bochenek, 2014; Cheaitou et al., 2019). For this reason, the Most Economically Advantageous Tender (MEAT) method has recently been receiving more attention from the researches and authorities, because considering multiple criteria in the public's procurement decision making process, such as quality, resources capacity, financial stability, experience in the market, innovation, sustainability and contribution to the society, in addition to the cost, can contribute to an increase in the efficiency of the public funds spending and thus to a lower probability of the issues related with the lowest-price method (Cheaitou et al., 2019; Grzyl et al., 2018).

With the increase of the relevance of Public Procurement, the open government data initiatives have also increased (Attard et al., 2015), reflecting one of the most promising policies, the e-procurement. This policy has gaining ground mainly in developed countries, in particular in European Union common market (SIGMA, 2016). In case of Portugal, a reformed was introduced in 2008 where it started to be mandatory the introduction of electronic public procurement information, which constitutes the repository of public contracts – portal BASE[1]. Hence, the introduction of the e-procurement has been mainly motivated to keep a greater transparency on political decision making (Bergman & Lundberg, 2013), standardization and digitalization of the contracts (García Rodríguez et al., 2019; Twizeyimana & Andersson, 2019). In addition to this, there are also other factors that driven the implementation of this policy, such as economic given that e-procurement permits a higher accuracy in the estimation of cost (Fry et al., 2016), risks (Bloomfield et al., 2019), bidders participation (Ballesteros-Pérez et al., 2016); social once it promotes less tolerance for inefficient political management or irregularities (Titl

[1] https://www.base.gov.pt/Base4/pt/

et al., 2018); and technological, as this policy allows to gather and archive a higher volume of data available associated with public procurement that can be used in advanced analytics studies (Varian, 2014), possible due to a rapid development of computational capacity, information management technologies and computer science.

Given all these factors, it is relevant to analyze public procurement data, as it can provide valuable insights to various stakeholders in both the private and public sectors. This analysis can facilitate the prediction of future behavior and enable informed operational decision-making (Jap & Naik, 2008). Hence, a valuable study involves predicting the number of public tenders won by companies and identifying the key factors that influence success in a multi-criteria context. Therefore, developing an accurate estimator would be highly beneficial for companies intending to participate in public tenders, as it would serve as a valuable tool for project management decision-making. It would aid in determining which tenders to pursue, whether to submit bids, and assessing the cost-value trade-offs associated with participating in specific public tenders. Furthermore, understanding the factors that impact the likelihood of winning a tender contributes to the formulation of effective tendering strategies and enhances self-assessment within the market. By evaluating the potential impact of bid adjustments on the chances of success, this analysis is particularly advantageous for small and medium-sized enterprises, which often face greater challenges during the assessment stage. From the perspective of public authorities, this estimation and factor determination process offers additional benefits. It helps characterize and identify the criteria that contribute to the success of bidders in the public tendering process, facilitating the development of more objective evaluation methods based on non-price criteria (Baranovsky et al., 2020) and creation of policies that promote fairness in the market.

Having this said, the present study is important in the sense that it complements the existing literature, by providing an updated look at the key factors influencing the success of bidders in public tenders. It achieves this by leveraging a comprehensive dataset comprising public procurement data from various sectors, combined with a diverse and extensive set of business information. This approach allows for the application of regression models that have not been previously explored in similar studies. As such, this research contributes to bridging gaps in knowledge and provides up-to-date insights into the determinants of success in public tenders. Hence, the findings of this study have implications for both government entities and companies. On one hand, it facilitates better management of public resources by guiding on the factors that contribute to successful tender outcomes. On the other hand, the study offers valuable information to companies, serving as a key tool in their project management decision-making. With the understanding of the factors that influence success in public tenders, companies can tailor their strategies, improve their competitiveness, and increase their chances of securing contracts. By addressing the identified factors and their impact on tender outcomes, this study contributes to the advancement of public procurement practices, fosters a competitive market environment, and supports the growth and sustainability of companies operating within the Portuguese context.

1.3. STUDY GOALS

The expanding volume of public procurement data necessitates the development of effective tools for knowledge extraction and prediction. This thesis aims to leverage public procurement data from Portugal to improve our understanding of the factors influencing the number of public tenders won by companies each year. Specifically, it seeks to demonstrate that factors beyond price, such as sector experience, public procurement experience, company characteristics, resources, innovation capacity, and financial stability, might impact a company's success in the procurement process. Moreover, this study seeks to contribute to the existing literature by employing a methodology not previously explored in the public procurement context. This involves analyzing public tender data across multiple sectors and incorporating a comprehensive set of business information variables not utilized in previous studies. Furthermore, considering that public procurements exhibit unique characteristics across different contract types, this study aims to investigate whether the identified factors and their impacts vary across contract types.

The analysis is structured into six sections. Section 2 presents a review of existing literature and relevant studies conducted by other researchers in this field. Section 3 provides a detailed description of the public procurement and company datasets, which are processed and merged to create a unified dataset. The public procurement data is sourced from the BASE portal, covering reported public procurements from 2008 to 2020. For the purpose of this study, only public contracts falling under the competitive procedure category are considered, as these tend to involve open bidding without invitation and typically involve higher contract values. Non-Portuguese companies are excluded from the analysis. The second dataset comprises information on Portuguese companies from 2011 to 2019, encompassing variables that characterize their experience, management, resources, and financial stability. Additionally, data quality rules are applied to ensure the inclusion of the most reliable and quantifiable public tenders. Section 3 also addresses limitations in the data, provides an exploratory analysis, and outlines the methodology employed.

Section 4 presents the empirical strategy, describing different count regression models, including Poisson, Negative Binomial, Zero-Inflated and Hurdle models. These models are evaluated to select the most suitable one based on the study objectives. Finally, in Section 5, a summary of the results and conclusions is presented, incorporating evaluation metrics introduced in the methodology section.

2. LITERATURE REVIEW

In this section the available literature will be reviewed in order to get to know what studies have already been proceeded in the field of Public Procurement and identify how our study can complement the work already published.

The literature related to public tenders is very extensive, in particular most of these studies are related to a specific sector, such as construction auctions. In the scope of predicting the success of bidders in the public procurement process, there are several studies that use different methods such as statistical techniques (Maciej Malara & Mariusz Mazurkiewicz, 2012), machine learning (García Rodríguez et al., 2020; Greenia et al., 2014), economic scoring formulas (Ballesteros-Pérez et al., 2013), Neuro-Fuzzy technique (Pamučar et al., 2022) and social network analysis to study the impact of inter-organizational relationships on the company's success in winning public procurement procedures (Sedita & Apa, 2015). These articles can be considered relevant to our study, since all have as final aim to create a decision-support model that allows companies to predict their performance in the public procurement market and therefore help to formulate an adequate strategy when bidding.

To predict the probability of a company winning in public procurement and identify the factors of success (Maciej Malara & Mariusz Mazurkiewicz, 2012), the author applies different statistics methods of which: binary logistic regression, discriminant analysis and cluster analysis, on Polish tenders' data. Results of this research show that bidding for a tender jointly with another contractor significantly increases the odds of winning the bid. Furthermore, the increase in the maximum price and an increase in the minimum price occurring in the tender also influence positively the changes of victory in the public tender, while the increase of the price proposed by the company implies a slight reduction in the chances of winning the tender. Although the methodology used is to predict a binary dependent variable and not a count dependent, this research is still relevant for our study since their method of variables selection and conclusions of the variables with higher impact on winning a public tender can be incorporated in our analysis.

Another relevant study is the Bidder Recommender for Public Auctions (García Rodríguez et al., 2020) which research aims to develop a pioneering algorithm to recommend potential bidders using machine learning methods. Therefore, this search engine recommends a group of companies for each tender, based on a framework of three sequential phases, that makes use of tender's information and companies' information from Spain. In the first phase, the authors train a Random Forest Classifier model to forecast the winning company of the tender. After predicting the winner of the tender, the algorithm gets the business information of the respective firm, to set the profile of the winning company. This profile is then used to set the filtering criteria to search a group of companies, in the companies' dataset, that has similarities criterion comparing to the winning company in terms of business, economic and technical characteristics, such as main business activities, location, human resources, size and financial performance. In the last phase, the recommender's results are evaluated, the evaluation metric selected is the Accuracy, i.e., the percentage of tenders where the winning company is within the recommended companies group. However, the data used in this research has some limitation on, given that the tender's dataset only identifies the winning company for each tender and not the rest of the bidders that bid on the tender, and for more recent tenders the availability of business information was lower, since the annual accounts of some winning companies were still not available. Regardless of these limitations, the final results of bidder recommender were positive, given the final accuracy obtained. Having this said, this literature can be considered as relevant to our study, not just because it aims to predict the winner company for each tender considering different types of tenders, contrary to other studies that usually are more focused in a specific industry, but also because it tries to set the profile of the winning company by merging the business information to the tender's data. Nevertheless, in this case the business data is only integrated with the tender's data after running the prediction model and is not used to train the predictive model as we aim to do in our study.

Outside the scope of public procurement, but also with the aim of predicting success in a bid, is the study performed in the IT Outsourcing market that develops a Win Prediction Model for IT Outsourcing Bids (Greenia et al., 2014), to identify the deals that are more likely to win and what factors, other than the price, correlate with the winning contracts. The author believes that this model can help the business to prioritize scare companies' resources to deals that are likely to be won and ensure that there are resources available to pursue the most attractive deals first. By understanding what factors influence the client's decision, the company can also change aspects of the deal that will raise the odds of winning the bid. The study concluded that factors like industrial sector, number of competitors, geographical region and deal complexity are some of the factors that most impact the odds of winning a contract.

To determine bidding behaviors and forecast the probability of obtaining a particular score and/or positions among competitors when bidding for capped tenders (Ballesteros-Pérez et al., 2013), i.e., tenders with a maximum price threshold set by the auctioneer, it was developed a Bid Tender Forecast Model that consists of iso-Score Curves Graph and Scoring Probability Graphs that the company can use to identify how the variation in the bid price can affect the company's score in the tender, and Position Probability Graph to study the likely position in the tendering results among competitors.

Another technique already studied is the Neuro-Fuzzy algorithm, to retrieve a rank prediction of an observed company compared to its competition when biding in a public tender in the construction market (Pamučar et al., 2022). This model was considered by the author, to successfully predict the public tender´s results and rank of the analyzed company, and hence able to provide the possibility of predicting the company's performance in the forthcoming business period as well as its position on the market, and thus allowing managers to adopt sustainable business strategies in the future. A limitation of this study is the fact that it only considers the variables related to the public tender´s and not with the business information in addition to the fact that only one company is considered. For this reason, the author suggests the identification of additional parameters in line with the specifics of the environment in which the company operates.

Furthermore, the literature on small and medium-sized enterprises (SMEs) in the context of public procurement has gained attention in recent years by the researchers. Data shows that SMEs are still underrepresented in public procurement (Glas & Eßig, 2018) and the high-value contracts awarded by these companies are usually fewer in contrast with a slightly more low-value contracts compared to the proportion of SMEs in the business sectors (Nicholas & Fruhmann, 2014). These statements motivated some authors to investigate the existence of equality in the public procurement market for SMEs, by predicting the size of the winning company in a public tender (Björkegren, 2022) and identifying the factors that most impacts the odds of a SME's success in public procurement (Glas &

Eßig, 2018). In both studies, the public procurement data was integrated with business information, such as turnover of the company and number of employees.

To predict the size of the winning company in a public tender (Björkegren, 2022), different models were applied to Swedish tenders' and companies' data, but the Support Vector Machine (SMV) was the model with better results. Furthermore, given that companies with different sizes have different characteristics, the authors decided to also evaluate the performance of all the models within each class. Results show that the SVM model performs better with micro, medium and big companies, while the Random Forest (RF) model performs better in the data of small companies.

Moreover, the identification of the main factors that influence the success of SMEs in public procurement was applied to German public procurement data together with the respective companies' data (Glas & Eßig, 2018). By applying a binary logistic regression in German public tenders' data, the study concludes that a higher number of lots *per se* do not reflects an increase in the chances of an SME's bidding success, as was expected by the author in the beginning of the literature. On the other side, the increase in factors like competitiveness in the award procedure and number of companies participating in the public tender (number of competitors) increase the likelihood that an SME will be successful in the bidding process, while the increase in the award volume of the contract decreases the likelihood of the SME's success in the public tender's bidding.

Adding to the described literature, there are other interesting articles that also aim to predict other factors in the context of the public procurement, such as the prediction of the number of bidders in a public tender and the respective determinants (Ballesteros-Pérez et al., 2016, 2019; Hattori, 2010; Stiti & Yape, 2019), prediction of the final price award of public procurement and the influencing factors (García Rodríguez et al., 2019; Grega & Nemec, 2015; Rodriguez et al., 2021; Stiti & Yape, 2019), identification of public procurement contracts with high risk of non-performance (Ovsyannikova & Domashova, 2020) and identification of possible fraud, corruption and collusion in public procurement activities (Lyra et al., 2021, 2022).

Table 2.1 shows an overview of variables used in the most relevant literature described previously. This overview is then divided into two parts, the Procurements Variables, which are the variables that characterize the public procurements and the Companies Variables, that are used by the authors to describe the companies that bid in the public procurements. To describe public procurements the most common variables used are the Nature of the Work (CPV), Procurement Price, Contracting Entity Location, Contracting Entity Name, Number of Competitors, Procurement Duration/ Deadline, Procurement Procedure Type and Lowest and Highest bid submitted in the procurement. When selecting companies' data, the authors consider as relevant the economic resources that can finance the project, like Operating Income, EBIT and EBITDA, the resources capacity, such as Number of Employees, the sector in which the company is specialized (NACE2 and NAICS) and the companies' geographical location. Other variables that would also be relevant to include are the variables that can measure the company solvency and its risk of failure or distress. Different studies regarding bankruptcy prediction consider as more relevant variables the Company Size, Company Age, Current Ratio, Solvency Ratio, Shareholders Liquidity Ratio, Cash Flow Return on Asset, Return on Activity, Return on Equity (Jones & Wang, 2019; Marilena & Alina, 2015; Yu et al., 2014).

In addition to these literatures, there are also several studies related to the selection of optimal bidder in public procurement (Bergman & Lundberg, 2013), that use different techniques such as the data envelopment analysis (Falagario et al., 2012), multicriteria decision making (Dotoli et al., 2020), competitiveness and position performance (Ballesteros-Pérez et al., 2015), and evaluation of criteria selection in the public procurement procedures (Baranovsky et al., 2020; Bochenek, 2014; Cheaitou et al., 2019; Grzyl et al., 2018; Lorentziadis, 2020).

In particular, the criteria selection evaluation shows the relevance of factors for contracts awards other than price, such as quality, expected implementation deadline, financial and technical capacity and the professional experience and knowledge of the suppliers. In Ukrainian public procurement the most important non-price criteria are analyzed and results show that factors such as payment term, distance, confirmation of experience years in the market, expected execution time, qualification of employees and confirmation of experience by the number of contracts and values of contracts are considered important in the tender evaluation process (Baranovsky et al., 2020). In addition to these, according to Article 53 of 2004/18/EC Directive for EU countries other factors such as technical merit, functional characteristics, environmental characteristics, and cost-effectiveness are also relevant when regarding contract awarded based on the most economically advantageous tender (Bochenek, 2014; Lorentziadis, 2020). Furthermore, other relevant criteria identified are the financial stability and management capabilities of the companies to measure the risk of bankruptcy, available resources, and capacity of innovation (Cheaitou et al., 2019).

In the context of using Data Science applied to the field of economics, there are also other interesting literatures that makes use of methodologies like Markov Chain model (Damásio et al., 2018; Damásio & Mendonça, 2019; Damásio & Nicolau, 2014, 2020), machine learning (Vaz, Bação, et al., 2021) and spatial analysis (Vaz, Cusimano, et al., 2021; Vaz, Damásio, et al., 2021).

Therefore, it's possible to observe that the number of literatures about public procurement is vast and that forecasting, and prediction techniques are widely studied in this field, with the vast majority being applied to construction contracts, given the bigger dimension of these projects. However, most of them presented several limitations given that many of the literatures either only analyze tender's data or in case of combining business data usually just one concrete company is considered, or the amount of data is small, or only tenders in a specific industry are reported. Therefore, the present study overcomes most of these limitations, since it considers a large dataset for both tender's data and company's data in all business sectors. In addition to this, the current study will also complement the presented literatures, in the sense that it will include more companies' variables that haven't been considered yet, mainly related to the financial stability measurement.

	Variables / Literature	(Maciej Malara & Mariusz Mazurkiewicz, 2012)	(Ballesteros-Pérez et al., 2013)	(Greenia et al., 2014)	(Sedita & Apa, 2015)	(Glas & Eßig, 2018)	(García Rodríguez et al., 2019)	(Stiti & Yape, 2019)	(García Rodríguez et al., 2020)	(Ovsyannikova & Domashova, 2020)	(Pamučar et al., 2022)	(Björkegren, 2022)
Procurement Variables	Contracting Entity Name						x		x			x
	Contracting Entity Type										x	x
	Contracting Entity Location			x			x	x	x			
	Procurement Date						x		x			
	Procurement Price		x			x	x		x	x		
	Effective Contractual Price						x			x		
	Duration or Procurement Deadline		x				x		x	x		
	Type of Processing[2]						x					
	Procurement Work Type[3]						x					
	Procurement Procedure Type[4]					x	x		x			
	Procurement Award Procedure Type[5]							x				
	Nature of Work or CPV		x				x	x	x	x	x	x
	Time for Preparation							x				
	Procurement Description Length											x
	Number of Procurement Publications											x
	Number of Award Criteria											x
	Price bid by the company	x						x				
	Lowest bid submitted in the Procurement	x	x					x				
	Highest bid submitted in the Procurement	x	x					x				
	Presence of Co-operators	x										
	Number of Competitors			x		x	x				x	
	Number of Lots					x						x
Company Variables	Company / Winner NIF						x		x	x		
	Company Size					x						x
	Last Available Year Info								x			
	Company Location								x			
	Company Status								x			
	Company Experience				x							
	Number of Employees				x	x			x			
	Operating Income				x	x			x			
	EBIT								x			
	EBITDA								x			
	NACE2 and NAICS			x					x			
	Liquidity Ratios									x		
	Profitability Ratios									x		

Table 1 - Variables considered in the analysed articles.

[2] Indicates the processing type of the procurement: ordinary, urgent, or emergency.
[3] Indicates the work type of the procurement: public works, services, goods and supplies.
[4] Procedure by which the contracts was awarded: open, restricted, negotiated with advertising, negotiated without publicity, competitive dialogue, internal rules, or others.
[5] Indicates whether a contract award criterion is the lowest price or MEAT.

3. DATA AND METHODOLOGY

In this section the elaborated process is described, from the data collection and transformation to the selection and application of the models. To achieve the goal of the study and determine which models to use in the methodology, first the data is collected and cleaned in to make use data of quality, then some exploratory analyses are done in order to better understand the data and its patterns and finally a careful variables selection is done based on the literature review and interpretation of statistic models that show how variables explain each other's. This will permit to get the final dataset that will be used in the selected models so that its results can be presented in the discussion section.

3.1. Data Collection

The present study will have in consideration two separate datasets. One dataset consists in Portuguese public procurement data extracted from the portal BASE through a web scrapping job developed in a previous study to scale behaviors of public procurement activity (Curado et al., 2021). The portal BASE is an open data platform managed by the IMPIC institution, and the data extracted from it follows the unified procurement format and contains the procurements celebrated between 2008 and 2020. To enrich the public procurement data, the second dataset is related to business data, and this is extracted from the ORBIS, which is a global company database that contains information on millions of private companies and entities across the globe. The company's dataset includes data between the years of 2011 and 2019, which consists mainly of company, managers, stakeholders and financial information. To achieve the goals of this study, both datasets will then be merged, and consider a timeframe between 2012 and 2020.

3.2. Data Description

The fields of Public Procurement dataset are explained in Appendix A. This data set consists of 32 variables that describe all the details regarding public procurements and comprises a total of 1.214.393 contracts. A remarkable limitation of this data is that for each procurement only one of the winners is identified in the cases where there is more than one awarded bidder. The Companies dataset characterizes 910.766 companies, that are possible contestants of the public procurements. This dataset contains 36 variables, which are divided into four sections, the general information, stakeholders' information, the managers' information, and financial measures, and these are described in Appendix B.

3.3. Data Cleaning and Transformation

Open data still has come challenges and risks related with data quality. For this reason, it is necessary to do a data cleaning of both public procurement dataset and companies' dataset, as this might contain possible human errors, such as wrong values, missing values and incorrect formatting that cannot be verified automatically. Hence, the identification of these issues will be identified and removed during the data cleaning stage. In addition to this, the data types uniformization and creation of new variables are done in the data transformation stage, to get the final dataset that will be used in the modeling stage.

3.3.1. Data Cleaning

In the data cleaning stage, the first step is to filter the data so that it goes according to the scope of this study, and it follows our data quality requirements. Thus, the following filters are applied:

1. Public Procurements Dataset:
    a. Contracting procedure type equals to ‘Public Tender’.
    b. The ‘Contracted’, ‘Contracting’ and ‘Contestants’ fields can’t be null.
    c. The ‘Contracted’ company must exist in the list of ‘Contestants’.
    d. The ‘Initial Contractual Price’ cannot be null and is higher than zero.
    e. In case the contract is finished it can’t be due to cancellation.
2. Companies Dataset:
    a. The company must have a Portuguese VAT number.
    b. The company must have data for all the financial fields.
    c. The company must have an ‘Operating Revenue’ higher or equal to zero.
    d. Select companies’ information before 2020, given that this year contains data with bad quality.

After applying the conditions for the inclusion of the observations, the data quality of the fields in both datasets are analyzed by assessing and treating the missing values. Appendix C and Appendix D show the percentage of missing values in the fields of the Public Procurement dataset and Companies dataset, respectively. Regarding contracts fields, some of the variables, such as ‘close_Date’ and ‘end_of_contract_type’ have a high number of nulls, however this is expected since these fields are only filled only when the contract is closed. For this reason, it was decided to keep these fields. On the other side the columns ‘contract_status’ and ‘cocontratantes’ have 100% of the fields null, so these fields will be discarded and no longer be considered in the current analysis. In the companies’ dataset, given the low percentage of missing values, these observations were simply removed from all the columns except for ‘NACE_Rev2’ and ‘LastAvail_FiscalYear’, since these are fields with lower importance. At the final of the cleaning stage, the public procurement contained a total of 335.207 records, with 47.410 public tenders and 12.903 contestants and the companies’ dataset contained a total of 6.829.781 records regarding 834.483 companies.

3.3.2. Data Transformation

After cleaning the data it’s important to make the data consistent, create new features, split the procurement data into the different contract types and combine the respective public tenders’ datasets with the business information, to generate the final datasets that will be used in the modeling stage. Hence the transformations steps consist of:

1. Selection of most relevant variables in public tender data and business data;
2. Convert fields to uniform scale and datatypes;
3. Create new variables in public tenders’ and companies’ datasets;
4. Create four different public procurement datasets: Consolidated, Public Works, Goods/Supplies, Services;
5. Aggregate each public tenders’ dataset;
6. Join each public tenders’ dataset with the companies’ dataset.

As a first step, an initial selection of the fields that can be considered as independent variables in the models is performed. This selection is made for the variables that are considered as possible factors of influence for the success of a company when bidding on public tenders and is based on the literature review realized in section 2, together with a critical analysis. Therefore, it is considered variables that are mentioned in previous articles, such as the ones regarding the contracts information and the variables that characterize the experience of the companies in the public tender procedures. In addition to these variables there are others not previously studied but considered as relevant to characterize the companies, such as their experience in the market, capacity of resources, financial stability, profitability, and efficiency.

After the selection of variables, the subsequent task involves establishing the scale and data types for each variable to maintain consistency across datasets and facilitate analysis. Consequently, in the public tenders' dataset, the variables 'initial_contractual_price' and 'total_effective_price' are transformed to the thousands scale to align with the amount variables in the company's dataset. Subsequently, appropriate data types are assigned to each variable based on their respective definitions. By ensuring uniform scales and defined data types, the datasets meet the necessary criteria for generating new variables that enhance the data. Appendix E and Appendix F provide a comprehensive listing and description of the newly created variables derived from the original inputs in the public tenders' and companies' datasets, respectively.

After all the variables are defined, four variations of the public tenders' dataset are created, one that includes all the contract types and another three for each contract type: Public Works, Goods/Supplies and Services. This is needed given that one of the goals in this study is to understand if the influencing factors in the success of a company in the public tender's process can vary based across industries. Thereafter, it is necessary to aggregate all the variables in the public tenders' datasets on Contestant NIF and Year level. Hence, with this aggregation the dependent variable is created, which will equal to the sum of the 'Won' column that indicates whether the Contestant won the tender or not.

Finally, having the public tenders and business information on the same level, it's possible to merge each of the public tenders' datasets with the companies' dataset. To perform this, the column 'Contestant_NIF' in the public tenders' dataset is merged with the 'Company_NIF' in the companies' dataset, and the column 'Year' of the public tenders' with the previous 'FiscalYear' available for the respective company. This means that for the financial information of the contestants, data from the previous year is considered, compared with the contract's year. It was decided to use the fiscal data from the previous year, given that in real life the data of the current year is usually only available at the end of the fiscal year. This step is crucial to be able to characterize the contestants and winners of the public tenders in the public procurement's dataset.

The final list of variables resulting from these transformations and respective descriptions are presented in Table 5. This table categorizes each of the variables in different groups: experience with public tenders' procedures, experience in the market, resources capacity, profitability, financial stability and efficiency. In addition to this, having performed all the steps described the new datasets are now comprised by:

- Consolidated – 27.794 records, 10.282 companies.
- Public Wors – 9216 records, 2840 companies.
- Goods and Supplies – 10.764 records, 4333 companies.
- Services – 12.341 records, 5595 companies.

Measure	Name	Description	Type
General Information	Contestant_NIF	Fiscal number of the company.	String
	Year	Year (N) in which we are going to analyse the number of won tenders.	Int
	FiscalYear	Fiscal year to which financial measures are respective to.	Int
Success in public tender	Won	Number of Public Tenders that company won in that year (Dependent Variable)	Int
	initial_contractual_price	Total amount of initial contractual price by the company in the Year N.	Float
	total_effective_price	Total amount of effective awarded price by the company in the Year N.	Float
Experience in Public Tenders	tender_closed	Number of closed tenders in the Year N.	Int
	Nbr_ParticipatedTenders	Number of Public Tenders to which the company has bid in the Year N.	Int
	ParticipatedTenders_Cum	Number of cumulative tenders to which the company bid since 2012.	Int
	Won_Cum	Number of cumulative tenders awarded by the company since 2012.	Int
	TendersClosed_Cum	Number of cumulative tenders that the company closed since 2012.	Int
	avg_execution_deadline	Average days of execution deadline for tenders awarded in the Year N.	Int
	avg_NbrContestans	Average number of contestants in the tenders participate the company.	Int
	avg_NrOfContractingEntities	Average number of entities involved in the awarded tenders.	Int
Performance in closed tenders	avg_ContractualPrice_Deviation_pct	Average deviation in percentage of the total effective price of the closed tender's vs the initial contractual price.	Float
	avg_ExecutionDays_Deviation_pct	Average deviation in percentage of the effective execution days of closed tenders vs the execution deadline.	Float
Management	Nbr_current_ directors_managers	Number of current directors and managers in year N.	Int
	Nbr_Shareholders	Number of shareholders in year N.	Int
Experience in the Industry	Company_Age	Number of years since the incorporation date of the contestant.	Int
	Company_Size	Company size at the Year N. List of Values: 4 – very large, 3 – large, 2 – medium, 1 – small.	Int
	NACE_Rev2	Classification of economic activities in the European Community.	String
Resources' Capacity	Total_Assets_EUR	Refers to the total amount of assets owned by the contestant.	Float
	Operating_Revenue_EUR	Refers to the income a company generates from its primary business activities.	Float
	Nbr_Employees	Total number of full-time employees of the company.	Int
	Shareholders_Funds_EUR	Refers to the amount of equity in a company, which belongs to the shareholders. The amount of shareholders' funds yields an approximation of theoretically how much the shareholders would receive if a business were to liquidate	Float
Financial stability	Current_Ratio	Is a liquidity ratio that measures a company's ability to pay short-term obligations or those due within one year. A current ratio that is lower than the industry average may indicate a higher risk of distress or default. And a company that has a very high current ratio compared with its peer group, it indicates that management may not be using its assets efficiently.	Float

	Solvency_Ratio_pct	It indicates whether a company’s cash flow is sufficient to meet its long-term liabilities and thus is a measure of its financial health. An unfavourable ratio can indicate some likelihood that a company will default on its debt obligations. A high solvency ratio is usually good as it means the company is usually in better long-term health compared to companies with lower solvency ratios.	Float
Profitability	PL_Before_Taxes_EUR	Profit before tax is a measure that looks at a company's profits before the company has to pay corporate income tax.	Float
	PL_After_Taxes_EUR	Profit After Tax refers to the amount that remains after a company has paid off all of its operating and non-operating expenses, other liabilities and taxes.	Float
	Cash_Flow_EUR	Refers to the net amount of cash and cash equivalents being transferred in and out of a company. Cash received represents inflows, while money spent represents outflows.	Float
	Profit_Marging_pct	Is a measure to evaluate the company's profitability. Expressed as a percentage, it represents the portion of a company’s sales revenue that it gets to keep as a profit, after subtracting all of its costs.	Float
Profitability / Efficiency	ROE_PL_Before_Taxes_pct	Return On Equity measures a corporation's profitability and how efficient it is in generating profits. The higher the ROE, the more efficient a company's management is at generating income and growth from its equity financing.	Float
	ROCE_PL_Before_Taxes_pct	The Return on capital employed (ROCE) refers to a financial ratio that can be used to assess a company's profitability and capital efficiency. A higher return on capital employed suggests a more efficient company, in terms of capital employment	Float
	Revenue_Per_Employee	It measures how much money each employee generates for the firm, so how efficiently a particular firm utilizes its employees. Ideally, a company wants the highest ratio of revenue per employee possible because a higher ratio indicates greater productivity.	Float
	Cash_ROA	It is an efficiency ratio that rates how well a company is managing assets. In other words, Cash ROA tells analysts how much each dollar of assets is generating in earnings. A high cash ROA ratio means the company earns more net income from $1 of assets than the average company, which is a sign of efficiency. A low cash ROA ratio means a company makes less net income per $1 of assets, which is a sign of inefficiency.	Float
	ROA_PL_After_Taxes_pct	The Return on Assets (ROA) refers to a financial ratio that indicates how profitable a company is in relation to its total assets, so it can be used to determine how efficiently a company uses its assets to generate a profit. A higher ROA means a company is more efficient and productive at managing its balance sheet to generate profits while a lower ROA indicates there is room for improvement.	Float

Table 2 - List and Description of Final Variables in the datasets

3.4. EXPLORATORY ANALYSIS

This section aims to analyze the final dataset to gain a better understanding of the data and to identify potential patterns or indicators among the tenders and contestants, which will be considered in the subsequent stage of result analysis. From the perspective of tenders, Figure 1 illustrates that the Goods and

Supplies public tenders have the highest participation rate among contestants (45%), while Public Works tenders have a larger number of unique contestants (44%), indicating greater contestant diversity for this contract type. Furthermore, the awareness of public tenders has significantly increased over the years, as depicted in both Figure 2 and Figure 3, with a rise in the number of contestants and the total effective price awarded by companies. Notably, the Goods and Supplies contract type experienced the most substantial growth since 2012.

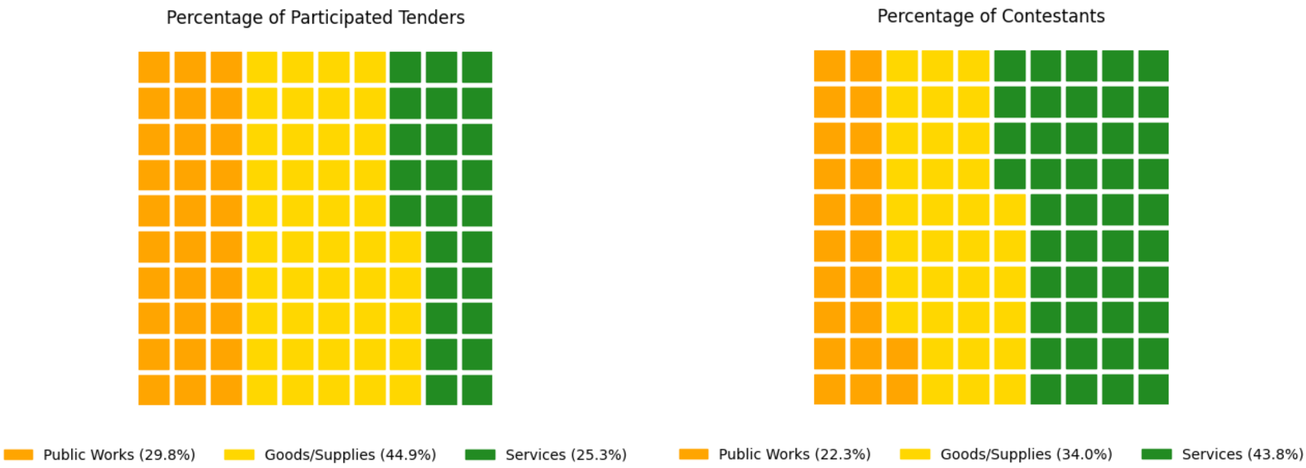


Figure 1 - Percentage of Participated Tenders and Contestants by Contract Type

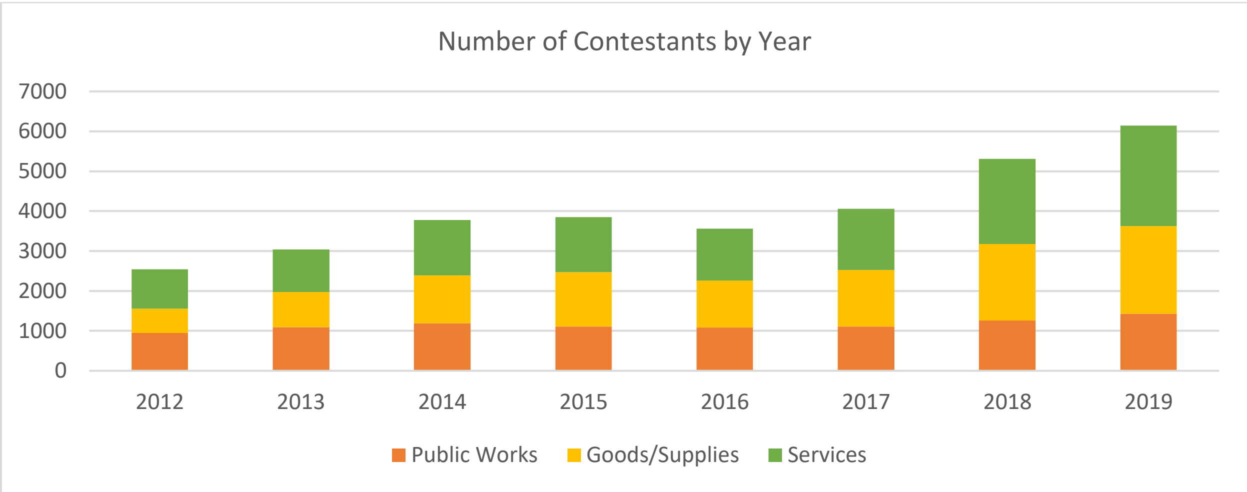


Figure 2 - Number of Contestants by Contract Type and Year

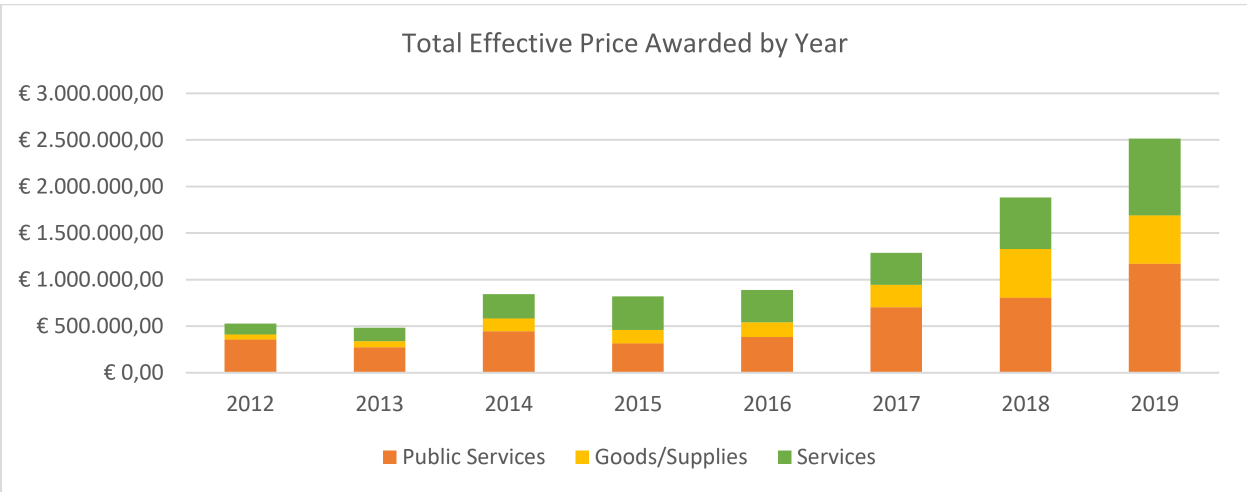


Figure 3 - Total Effective Price (Th) Awarded by Year

By analyzing the descriptive statistics of the four public tenders' datasets, distinctive characteristics can be identified among different contract types. The complete descriptive statistics for each dataset can be found in Appendix G, H, I, and J. Based on the information presented in Table 3 to Table 6, it is evident that Public Works tenders are distinguished by their larger scale in terms of Initial Contractual Price and Effective Price Awarded, despite having comparatively shorter execution deadlines compared to other contract types. The greater magnitude of these contracts may explain the higher number of participants, indicating a higher level of competition. In contrast, Goods/Supplies and Services contracts exhibit lower values for Initial Contractual Price and Effective Price Awarded, while being characterized by medium to long-term projects with significant variations in execution deadlines. Consequently, these contract types tend to attract fewer contestants.

Dataset	Consolidated	Public Works	Goods/Supplies	Services
count	27763	9216	10755	12314
mean	334,86	483,25	173,99	241,34
std	1858,55	2039,75	1129,18	1668,17
min	0	0	0	0
25%	0	0	0	0
50%	0	0	2,42	0
75%	139,03	261,48	91,84	90,42
max	82472,43	82063,7	49859,08	61599,14

Table 3 - Initial Contractual Price Descriptive Statistics by Contract Type

Dataset	Consolidated	Public Works	Goods/Supplies	Services
count	27763	9216	10755	12314
mean	333,06	482,99	171,82	239,37
std	1854,21	2042,15	1125,99	1660,27
min	0	0	0	0
25%	0	0	0	0
50%	0	0	2	0
75%	136,69	260,23	89,92	89,9
max	82472,43	82063,7	49925,71	61599,14

Table 4 - Total Effective Price Descriptive Statistics by Contract Type

Dataset	Consolidated	Public Works	Goods/Supplies	Services
count	13261	3586	5587	5283
mean	368,3	202,37	323,72	561,84
std	1433,63	222,61	1891,56	2146,33
min	1	8	1	1
25%	120	101	60	223
50%	262	162,5	221	365
75%	405	258	365	730
max	148432	9125	139000	148432

Table 5 - Execution Deadline Descriptive Statistics by Contract Type

Dataset	Consolidated	Public Works	Goods/Supplies	Services
count	27763	9216	10755	12314
mean	9,77	12,38	8,03	8,56
std	6,9	6,36	5,72	7,16
min	1	1	1	1
25%	5	8	4	4
50%	8	11	6	7
75%	13	16	11	11
max	66	49	45	66

Table 6 - Number of Contestants Descriptive Statistics by Contract Type

Another interesting analysis involves examining the impact of public tenders on company performance. Figure 4 presents the Total Amount Awarded by the companies in our dataset compared with their respective Total Operational Revenue. The Figure shows that in the earliest years, the expenditure on public tenders by the state from 2012 to 2014 exhibited a relatively lower magnitude, possibly attributable to the influence of the Troika's presence in Portugal during that time (Martins & Damásio, 2020). Furthermore, the figure demonstrates a consistent trend over time, revealing a positive association of 28% between the awarded price and operational revenue. Specifically, as the total amount awarded by the companies increases, there is a corresponding rise in operational revenue.

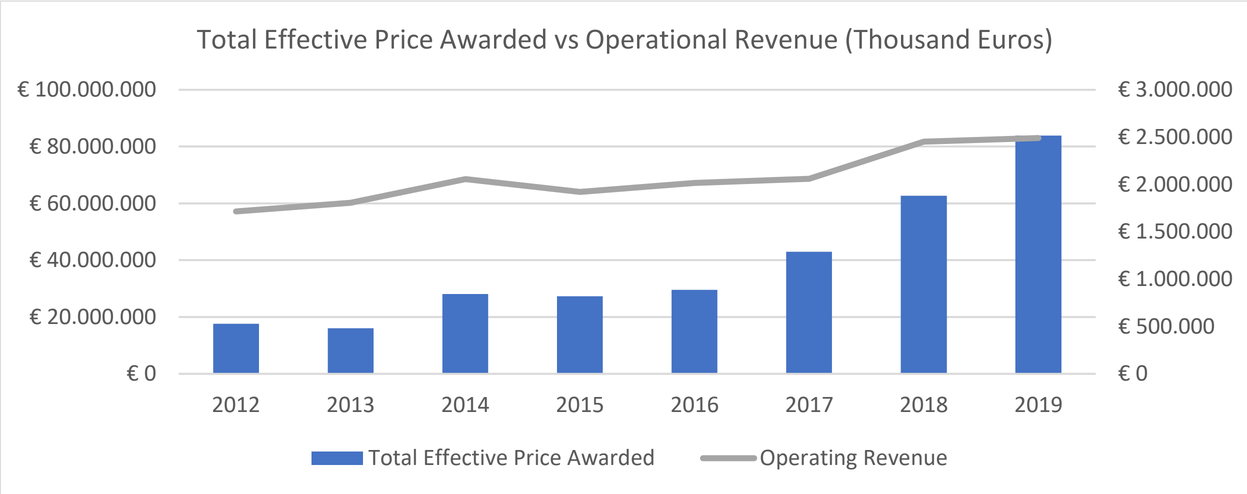


Figure 4 - Total Effective Price Awarded (Th) and Operational Revenue (Th) of the Contestants by Year

The subsequent analysis aims to examine the relationship between company size and the characteristics of participated and awarded tenders. Table 7 provides an overview of the dataset, revealing that the majority of represented companies are small and medium-sized across all contract types. Moving on to Table 8, it becomes apparent that larger companies tend to participate in a greater number of tenders compared to smaller companies, likely due to their higher human resource capacity. When considering different contract types, it is observed that small and medium companies predominantly participate in Goods and Supplies tenders, while large companies focus on public works tenders, and very large companies engage in services tenders. This might then present a challenge for smaller companies seeking to compete in these contract types.

When analyzing the total effective price and execution deadline of awarded tenders by the contestants, as depicted in Table 9 and Table 10 respectively, it is evident that both metrics increase as company size increases. In other words, larger companies tend to secure tenders with higher monetary value and longer durations compared to smaller companies. This can be attributed to the larger companies' greater resources and capabilities.

Lastly, Table 11 presents the average winning rate of companies overall and across contract types. This measure represents the proportion of successfully won public tenders out of the total number of participated tenders, serving as an indicator of contestants' success. Consistently, larger companies exhibit higher winning rates, ranging between 23.4% and 24.5%. This outcome can be attributed to the extensive experience and established reputation of large and very large companies in the market. When looking across contract types, it is notable that winning rates are typically higher in the Services category, while Public Works tenders yield lower winning rates. This discrepancy can be attributed to the larger scale and higher competition typically associated with Public Works projects.

Company Size	Consolidated	Public Works	Goods/Supplies	Services
1. Small	5.339	1.321	1.738	2.969
2. Medium	4.740	1.669	2.199	2.264
3. Large	1.223	296	658	698
4. Very Large	275	49	171	203

Table 7 - Number of Contestants by Company Size and Contract Type

Company Size	Consolidated	Public Works	Goods/Supplies	Services
1. Small	4,54	3,79	6,48	3,02
2. Medium	9,20	8,33	10,51	4,34
3. Large	15,41	16,50	14,60	7,60
4. Very Large	22,08	13,33	15,03	17,64

Table 8 - Average Number of Participated Tenders

Company Size	Consolidated	Public Works	Goods/Supplies	Services
1. Small	€ 51,37	€ 65,60	€ 36,01	€ 44,68
2. Medium	€ 246,97	€ 396,13	€ 103,21	€ 123,60
3. Large	€ 743,22	€ 1.589,19	€ 249,36	€ 392,52
4. Very Large	€ 2.462,29	€ 2.913,69	€ 1.241,70	€ 1.948,08

Table 9 - Average Total Effective Price Awarded (Th) by Company Size and Contract Type

Company Size	Consolidated	Public Works	Goods/Supplies	Services
1. Small	348	176	245	496
2. Medium	348	204	323	582
3. Large	394	223	343	570
4. Very Large	565	262	515	687

Table 10 - Average Execution Deadline by Company Size and Contract Size

Company Size	Consolidated	Public Works	Goods/Supplies	Services
1. Small	20,03	12,27	19,54	23,89
2. Medium	19,85	13,17	23,05	23,78
3. Large	23,42	10,88	26,11	26,85
4. Very Large	24,51	8,81	28,28	27,29

Table 11 - Average Winning Rate by Company Size and Contract Type[6]

To conclude the exploratory analysis session, a scatter plot analysis is conducted to examine the relationship between various independent variables that characterize the contestants and the dependent variable 'Won'. The objective of this analysis is to gain insights into how the contestants' characteristics influence their success in winning public tenders, specifically, the number of contracts won. Hence, Figure 5 displays the dependent variables that exhibit a more substantial relationship with the independent variable, providing insights into the characteristics of the contestants in terms of experience (Company Age, Cumulative Number of Participated Tenders, and Cumulative Number of Closed Tenders), resource capacity (Number of Employees and Operational Revenue), and financial performance (Profit Margin % and Solvency Ratio %).

Examining the three variables that depict the experience of the companies, a consistent pattern emerges where higher values in company age, cumulative number of participated tenders, and cumulative number of closed tenders correspond to a greater number of won tenders. Notably, in the Public Works dataset, the trend lines for these three variables are less pronounced compared to the datasets of other contract types,

[6] Winning Rate equals to the total number of won tenders divided by the total number of participated tenders by the contestant.

suggesting that experience may not hold the same significance in this specific contract type as it does in others.

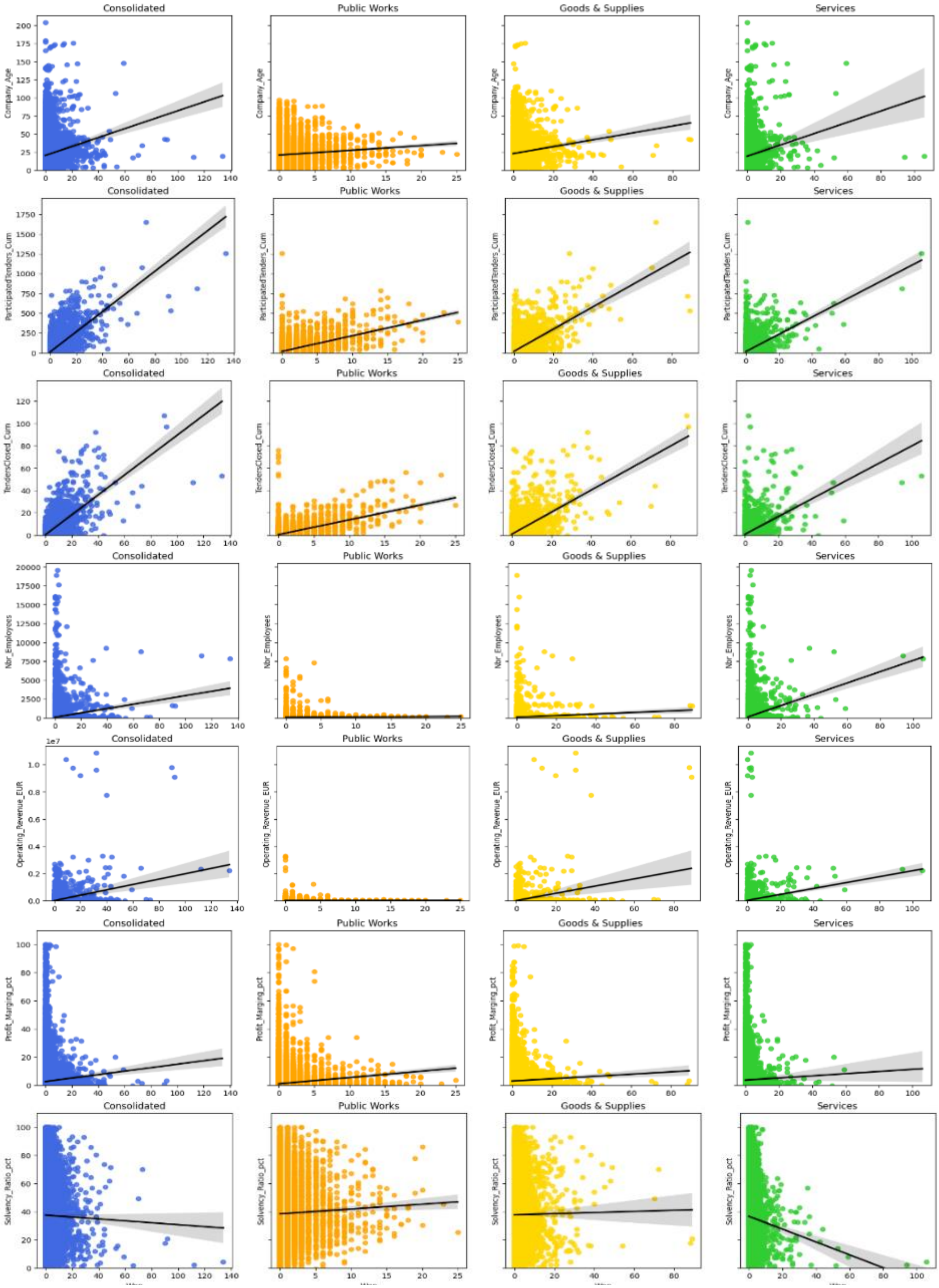


Figure 5 - Relationship between Contestant's Characteristics and the Won Variables

Regarding resource capacity, there is a positive impact on the number of won tenders, particularly evident in the Services contract types. The trend line for the number of employees is more pronounced in these contract types, which aligns with the notion that services often rely heavily on human resources. Additionally, the Operational Revenue, representing the income available for project financing, demonstrates a significant relationship with the independent variable, particularly in Goods and Supplies contracts. Interestingly, neither of these two variables appears to have a notable impact on the number of public tenders awarded in the Public Works category.

In terms of financial performance of the contestants, as expected the profit margin has a positive relation with the dependent variable, meaning that the fact that the contestants have lower costs compared it income might mean a good performance in terms of project and cost management and also a higher income available to finance the projects. On the other hand, the solvency ratio trend line, in general, declines a bit with the increase of the solvency ratio, however this varies across contract types. Hence, for Public Works and Goods and Supplies contracts, a higher solvency ratio tends to have a positive impact on the number of won contracts. This observation is particularly meaningful for Public Works contracts since a higher solvency ratio signifies a company's greater capacity to meet long-term liabilities and indicates better financial stability. Given that public works projects typically involve longer execution deadlines, contracting entities are inclined to select companies with a lower likelihood of defaulting during project execution. Conversely, the behavior of the solvency ratio in the Services contract category differs. In this industry, a lower solvency ratio is generally associated with a higher number of won tenders. This phenomenon can be attributed to several factors. A lower solvency ratio may indicate a high volume of total assets, providing more resources available to execute the required services. It could also signify that shareholders' funds are lower because the company has opted to leverage external funding sources, such as debt or equity, to finance expansion projects and growth initiatives.

In conclusion, the exploratory analysis conducted in this study revealed variations in the characteristics of public tenders across different contract types. Furthermore, it was observed that contestant characteristics have distinct impacts on each contract type. These findings reaffirm the importance of applying appropriate models and conducting separate analyses on distinct datasets to accurately identify the key influencing factors contributing to companies' success when participating in public contract bidding within specific industries. By recognizing the unique dynamics and requirements of each contract type, policymakers, contracting entities, and companies can make more informed decisions and strategies to enhance their performance in the public procurement process.

3.5. MODELING

To forecast the companies' success in the public tenders' procedures this study aims to predict the count of won public tenders by the companies per year and measure the impact of the different independent variables in this prediction. Hence, given that the dependent variable analyzed in this study represents a discrete variable obtained with a count of an event that is non-negative valued, a methodology that can produce convincing results is the Count Data Regression models. These models deviate from the assumptions of classical regression models, which rely on normality, linearity, and homoscedasticity. When dealing with a count response variable, these assumptions can be compromised, leading to biased, unstable, and less generalizable results (Yildirim et al., 2022).

Various alternative models have been proposed to better accommodate count data, including those based on Poisson and Negative Binomial distributions. Poisson regression (PR) and Negative Binomial regression

(NBR) models have gained significant interest across diverse fields such as medicine, biostatistics, biology, finance, demography, astronomy, business and management, earth sciences, social sciences, communication, and insurance. Hence there are key benefits of employing count data regression models of which we can highlight the enhanced interpretability, ease of implementation, good performance, and broad applicability (Chatfield & Zidek, 2002; Samprit Chatterjee & Ali S. Hadi, 2015). In particular for this study, the fact that count data regression models provide interpretable results is a relevant benefit, given that this provides a clear understanding of the relationship between the independent variables and the count outcome. Hence, the estimated coefficients can be directly interpreted as the effect on the expected count for a one-unit change in the corresponding predictor, holding other variables constant.

In Figure 6, it is presented the distribution of the independent variable 'Won_Tenders' in the Consolidated dataset. The same distribution for the Public Works, Goods and Supplies and Services, follow a very similar pattern and these can be found in Appendix K. Hence, upon examination of the independent variable's distribution on the datasets, it was observed that there is a significant left skewness primarily caused by an excessive number of zero values, indicating the no-occurrence in the variable of interest. This excess of zeros contributes to the problem of overdispersion, leading to an inflated sample variance (Cameron & Trivedi, 1998). Consequently, the variance exceeds the mean, contradicting the assumption of their equality. That's why in such a situation, classical PR and NBR models may not be sufficient for count data modelling. Although the effect of overdispersion over classical Poisson regression is more severe than Negative Binomial regression, which means that it can even be used to a certain extent in the existence of excessive zeros. For this reason, it is purposed to use in this study the models Negative Binomial Regression (NB), Zero Inflated Poisson Regression (ZIP), Zero Inflated Negative Binomial Regression (ZINB), Hurdle Poisson Regression (HP) and Hurdle Negative Binomial Regression (HNB). The model Poisson (P) will also be analyzed just for sake of comparison.

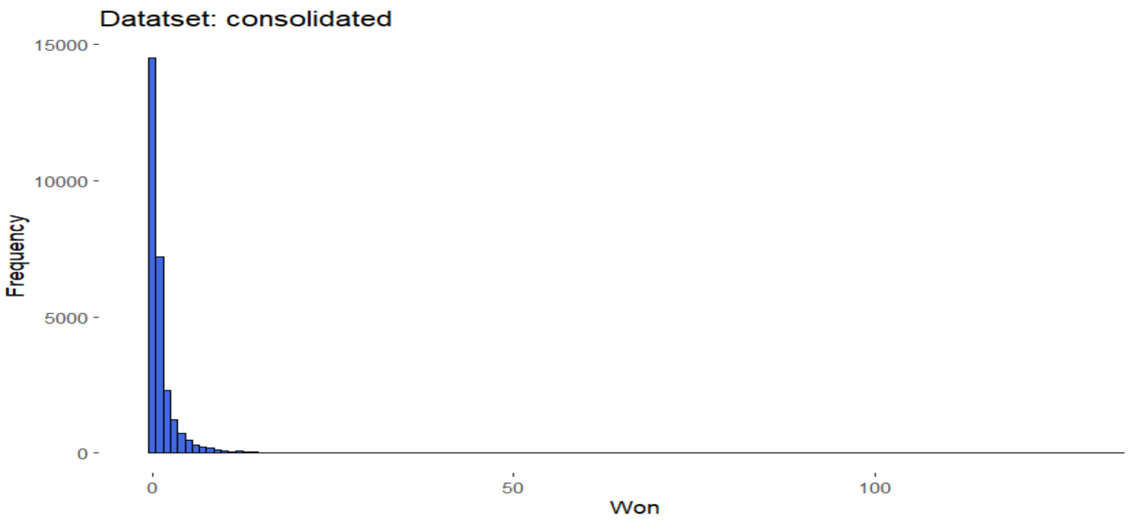


Figure 6 - Histogram with Distribution of Independent Variable 'Won' in Consolidated Dataset

Prior to applying the proposed models, is important to evaluate potential collinearity and multicollinearity issues within the dataset. Collinearity refers to a high degree of correlation or linear dependence between two or more independent variables in a regression model, which can lead to unstable parameter estimates, inflated standard errors, and unreliable interpretations of the relationships between variables (James et al., 2013). On the other hand, multicollinearity is a specific form of collinearity that occurs when there is a high degree of correlation among three or more independent variables in a regression model and which can also bring challenges since it affects the interpretation and precision of regression coefficients (Agresti, 2019). Hence, the presence of collinearity or multicollinearity in a count data regression model can cause problems in statistical inference. To address collinearity or multicollinearity, several techniques can be employed. The collinearity can be identified by analyzing the correlation matrix of the predictors, where a large absolute value in this matrix indicates a pair of highly correlated variables, and therefore a collinearity problem in the data. To analyze multicollinearity, the same approach cannot be used, given that is possible to exist multicollinearity between three or more variables even if no pair of variables has a particularly high correlation. Hence, the Variance Inflation Factor approach is used instead (James et al., 2013).

To analyze the correlation between the independent variables, it was decided to use the Spearman Correlation. This type of correlation is a non-parametric measure of the strength and direction of the monotonic relationship between two variables. Hence, it is used to assess the association between variables that may not have a linear relationship. This is particularly useful for our data, given that it does not follow a normal distribution. Unlike Pearson correlation, which measures the strength and direction of a linear relationship between variables, Spearman correlation is based on the ranks of the values rather than the actual values themselves. The Spearman correlation coefficient ranges from -1 to +1, where -1 indicates a perfect negative monotonic relationship, +1 indicates a perfect positive monotonic relationship, and 0 indicates no monotonic relationship. In Figure 7 is possible to observe the Spearman correlation matrix for the Consolidated dataset. The same analysis for the other datasets can be found in Appendix L. Therefore, is possible to observe that the following variables have a strong relation between each other: Nbr_ParticipatedTenders and ParticipatedTenders_Cum (0,82), Won and Won_Cum (0,79), Initial_Contractual_Price and Total_Effective_Price (1), PL_Before_Taxes_EUR and PL_After_Taxes_EUR (0,95) and Cash_ROA and ROA_PL_After_Taxes_EUR (0,87). In addition to this, is also possible to observe that other variables also have a high correlation, such as, the dependent variable Won with Total_Effective_Price, Initial_Contractual_Price, Tenders_Closed and TendersClosed_Cum variables; the

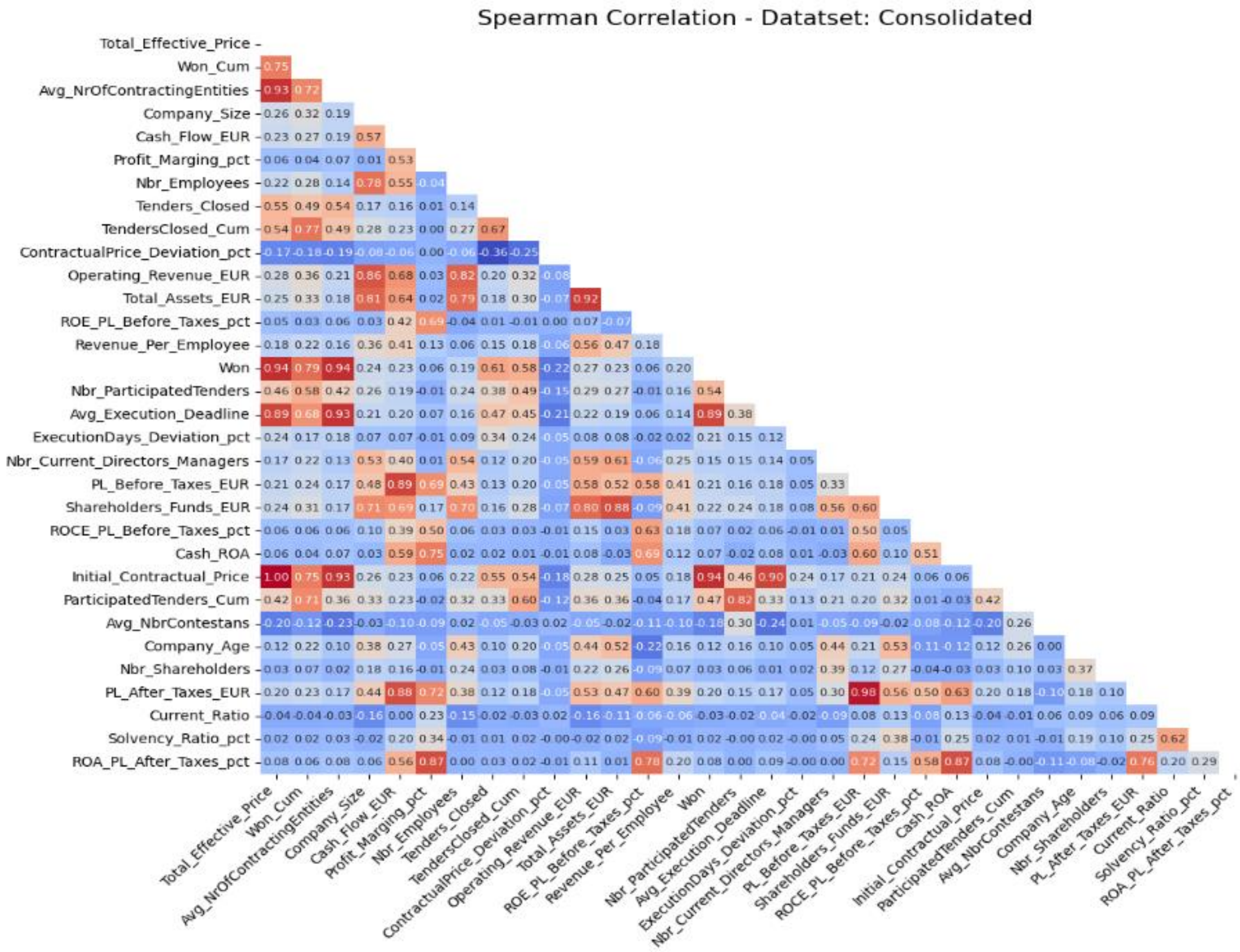


Figure 7 - Spearman Correlation of Consolidated Dataset

Profit_Marging_pct with Cash_ROA, PL_Before_Taxes_EUR, PL_After_Taxes_EUR and ROA_PL_After_Taxes_pct; the Operational_Revenue_EUR with Cash_Flow_EUR and Revenue_Per_Employee; the Shareholders_Funds_EUR with the Total_Assets_EUR and finally the Nbr_Shareholders with the Nbr_Current_Directors_Managers. The same variables also present a high correlation among each other in the other datasets, and to avoid redundant variables in the model, the following fields were removed from all datasets: Total_Effective_Price, Initial_Contractual_Price, Nbr_ParticipatedTenders, Won_Cum, Tenders_Closed, TendersClosed_Cum, Cash_Flow_EUR, Revenue_Per_Employee, PL_Before_Taxes_EUR, PL_After_Taxes_EUR, Shareholders_Funds_EUR, Nbr_Shareholders, Cash_ROA and ROA_PL_After_Taxes_pct.

As mentioned previously, to assess multicollinearity in the data a Variance Inflation Factor (VIF) is performed. Thus, the VIF quantifies how much the variance of the estimated regression coefficients is increased due to multicollinearity by measuring the level to which a predictor variable can be linearly predicted from the other predictor variables in the model. The formula for calculating the VIF of a predictor variable is given by: $VIF = 1/(1 - R^2)$, where R^2 is the coefficient of determination from regressing the predictor variable against all other predictor variables in the model. A VIF value of 1 indicates no multicollinearity, while VIF values greater than 1 suggest increasing levels of multicollinearity (James et al., 2013). To perform this analysis, it was used a poisson regression having 'Won' as the dependent variable. Then, a VIF score threshold of 10 was defined, given that VIF value above this threshold can already be considered problematic as it might significantly influence the stability of the parameter estimates (James et al., 2013). Figure 8 presents the results obtained in the Consolidated dataset, taking already into consideration the removed fields in the previous step. The same results for the remaining datasets are presented in Appendix M. By observing the VIF values in all datasets is possible to conclude that any of the remaining variables existing in the dataset after correlation analysis present a VIF score higher than the defined threshold, meaning that any independent variable is excluded in this step.

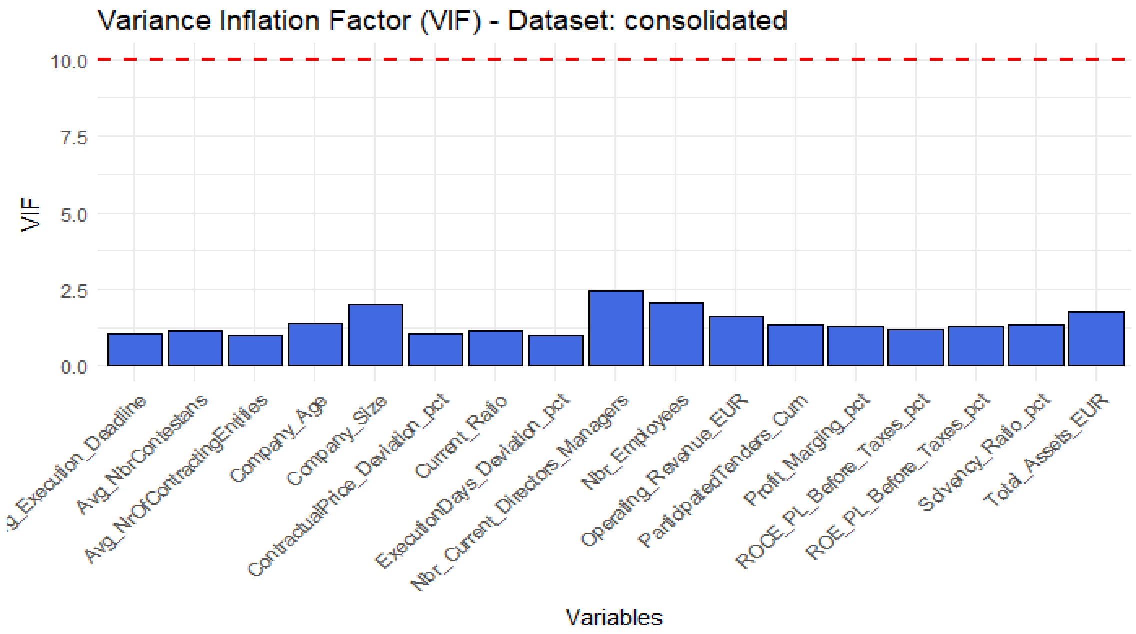


Figure 8 - Variance Inflation Factor of Dependent Variables in Consolidated Dataset

Before models are applied in the final data, the last analysis to perform is the presence of outliers in the dataset's variables. Count data regression models are typically resilient to moderate outliers, but the presence of extreme outliers can adversely affect the accuracy of model estimates and subsequently influence the results. Hence in Appendix N, O, P and Q there are presented the results from the boxplot analysis for the variables in the Consolidated, Public Works, Goods and Supplies and Services datasets, respectively. In the boxplots for all datasets, it's possible to see a clear presence of outliers that might influence the final results of the models, mainly in the variables like 'Avg_Execution_Deadline', 'ExecutionDays_Deviation_pct', 'Nbr_Current_Directors_Managers', 'Avg_NbrOfContractingEntities', 'Operating_Revenue_EUR' and 'Total_Assets_EUR'. Hence, to alleviate the potential influence of outliers on model estimates, some outliers are removed from the variables where these were more significant, such as Operational_Revenue_EUR and Total_Assets_EUR.

Following this step, to make sure that all variables have equal influence on the analysis a standardization technique is applied to the independent variables, to put the variables on the same scale. Given that the data in this study is skewed and non-normal, a Robust Scaling is applied, so that is possible to preserve the ranking and relative order of the data. Unlike other scaling methods, Robust Scaling uses robust statistics, such as median and interquartile range, that ensure that different variables are on a similar scale and resistant to extreme values, turning the data more robust to outliers. This technique is applied by subtracting the median and dividing it by the median absolute deviation (MAD) or the interquartile range (IQR).

Finally, to predict the response variable 'Won', the independent variables used in the models are: CompanySize, NbrEmployees, ROE_PL_Before_Taxes_pct, Nbr_Current_Directors_Managers, ROCE_PL_Before_Taxes_pct, ParticipatedTenders_Cum, Company_Age, Current_Ratio, Solvency_Ratio_pct , Total_Assets_EUR , Operating_Revenue_EUR and Profit_Marging_pct. The remaining variables Avg_NrOfContractingEntities, ContractualPrice_Deviation_pct, Avg_Execution_Deadline, ExecutionDays_Deviation_pct and Avg_NbrContestans were not considered, given that models wouldn't converge with the presence of these.

The theoretical background of the models used in the modeling stage is explained in detail in Section 4.

4. EMPIRICAL STRATEGY

This section describes the necessary components to achieve our study goals. First, the models used in the study are presented, followed by techniques to evaluate both model's goodness of fit and accuracy, so that the most suitable methodology is selected to achieve the study goals presented.

4.1. COUNT DATA REGRESSION MODELS

To model independent variables that result from the count of events, i.e., a response variable y that has non-negative whole integers, often the Poisson and Negative Binomial models are used. However, when the response variable presents an overdispersion due to an excess of zeros, these traditional models used for a common count distribution may not provide a reasonable fit for both zero and non-zero counts (Perumean-Chaney et al., 2013). For this reason, Zero-Inflated (Lambert, 1992) and Hurdle models (Heilbron, 1994; Mullahy, 1986) have been developed to model zero-inflation when the regular count models are unrealistic. These models have gained popularity in various fields, including public health services research (B. Neelon et al., 2012, 2016; B. H. Neelon et al., 2010), substance abuse (B. Neelon et al., 2012), occupational injury (Yau & Lee, 2001), medicine (Rose et al., 2006), public health (Cheung, 2002), ecological and environmental studies (Deepak K. Agarwal et al., 2002).

4.1.1. Poisson and Negative Binomial Regression Models

Poisson regression models have been widely used and popular in various fields due to its advantages like the interpretability and suitability on inference count data. The conventional Poisson models are based on poisson distribution which is given by:

$$\mathrm{P}(y_i, x_i) = \frac{e^{-\mu_i}\, {\mu_i}^{y_i}}{y_i!}, y = 0, 1, 2, \dots,$$

(1)

where μ_i is the mean parameter define by $E[\, y_i |\, \mathrm{x}_i] = \mu_i = \exp\,({x_i}'\beta)$, x_i corresponds to the vector of covariates, β is a $(k+1)$ x 1 parameter vector and y_i is the response variable. Poisson log likelihood is defined as follows:

$$L(\mu, y) = \sum_{i=1}^{n} \{\, y_i\, ln\,(\mu_i) - \mu_i - ln\,(y_i!)\}$$

(2)

Poisson distribution has conditions of independent events and equality of mean and variance (Consul & Famoye, 1992). Due to the limitation of equal mean and variance, the Poisson distribution is not appropriate to fit the count response, when the variance of the counts is much larger than their mean, i.e., when there is overdispersion in the data. A solution for this overdispersion can be solved by applying a gamma-Poisson mixture distribution, which is known as the *Negative Binomial (NB)* distribution. Hence NB regression models is a powerful tool which is highly effective on a wide range of count data applications. Therefore, unlike the conventional Poisson model, the convention NB model can deal with over-dispersion problem by adding an additional term into probability function and this function is defined as follows:

$$P(\, y_i, \mu_i, \tau) = \frac{\Gamma(\tau^{-1} + y_i)}{\Gamma(\tau^{-1})\Gamma(y_i+1)} \left(\frac{\tau^{-1}}{\tau^{-1} + \mu_i}\right)^{\tau^{-1}} \left(\frac{\mu_i}{\mu_i + \tau^{-1}}\right)^{y_i},$$

(3)

where the mean and variance of the negative binomial distribution are $E[y_i] = \mu_i\tau$ and $Var[y_i] = \mu_i\tau\,(1 + \mu_i)$, y_i is the response variable and τ is the dispersion parameter, if τ = 0, negative binomial approaches to the Poisson model (Hilbe, 2011).

A critical problem for these models' performance is the overdispersion due to excessive zeros in the response variable, which is not a rare situation in practical real applications. Therefore, the positive count values can be estimated more accurately via the conventional models, Poisson and NB distributions, but the zero counts cause zero-inflation on the model. To overcome the overdispersion issue, Zero-Inflated and Hurdle regression models have been proposed. These models account for excessive zeros separately in the estimation process. The modelling stage consists of two parts, one for estimation of positive counts and the other one for estimation of zero count values. Generally, while logistic or probit models are used to estimate zero values in a binary logic, classical poisson or negative binomial are used for positive count values (Yildirim et al., 2022).

4.1.2. Zero-Inflated Models

In a zero-inflated model (ZI), there are two-component mixture models combining a point mass at zero with a count distribution such as Poisson, Geometric or Negative Binomial (Lambert, 1992). The point mass at zero accounts for excess zeros in the data, while the count distribution characterizes the non-zero counts. Thus, there are two sources of zeros: zeros that come from the point mass and zeros that come from the estimation process of the count values. Hence, it is important to distinguish between two sources of zeros in a ZI model: "structural" and "sampling" zeros. Structural zeros arise from the inherent nature of the phenomenon being studied, while sampling zeros result from the estimation process of the count values. The structural zeros are represented by the point mass component, while the sampling zeros are assumed to follow a standard count distribution (e.g., Poisson or Negative Binomial) and which are assumed that were occurred by chance. The general structure of a ZI model is given as:

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1 - \pi_i)p(y_i = 0; \mu_i) & \text{if } y_i = 0, \\ (1 - \pi_i)p(y_i; \mu_i) & \text{if } y_i > 0, \end{cases} \tag{4}$$

This equation consists of a degenerate distribution at zero and an untruncated count distribution with a vector of parameters μ_i. If the count distribution follows a Poisson distribution, the *Zero Inflated Poisson model (ZIP)* is given by:

$$P(y_i, \mu_i) = \begin{cases} \pi_i + (1 - \pi_i)e^{-\mu_i} & y_i = 0, 0 \le \pi_i \le 1 \\ (1 - \pi_i)z\dfrac{e^{-\mu_i}\,\mu_i^{y_i}}{y_i!} & y_i > 0, \end{cases} \tag{5}$$

, where μ_i is the mean of the standard Poisson distribution. ZI model is also formulated as a latent variable model with an unobserved Bernoulli random variable z_i (Lambert, 1992):

$$z_i = \begin{cases} 1 \text{ if } y_i \text{ is stuctural zero,} \\ 0 \text{ if } y_i \sim \text{Poisson } (\mu_i), \end{cases} \tag{6}$$

,which establishes:

$$E(y_i) = E\left(E(y_i|z_i)\right) = (1 - \pi_i)\mu_i$$

(7)

$$Var(y_i) = E(Var(y_i|z_i)) + Var(E(y_i|z_i)) = (1 - \pi_i)\mu_i\,(1 + \mu_i\pi_i)$$

(8)

In the context of modeling the count component in a Zero-Inflated (ZI) model, Poisson regression assumes that the conditional mean is equal to the conditional variance. However, this assumption may not hold in certain situations. If the data exhibit greater conditional variance than expected in a Poisson model, it indicates overdispersion, which can arise due to factors such as heterogeneity in the data, the omission of important covariates in the model, or the presence of outliers (Cox, 1983; Dean, 1992; Payne et al., 2017). To address overdispersion in count data, an alternative approach known as Negative Binomial (NB) regression can be employed. By incorporating the NB distribution, the model accounts for the excess variation present in the data, providing a more accurate representation of the conditional mean-variance relationship. In the case of a Zero-Inflated Negative Binomial (ZINB) model, both excess zeros and overdispersion are simultaneously addressed, by combining the zero-inflation component with the NB count component. The *Zero Inflated Negative Binomial (ZINB)* model is then given by:

$$P(y_i, \mu_i, \tau) = \begin{cases} \pi_i + (1 - \pi_i)(1 + \tau\mu_i)^{-\tau^{-1}} & 0 < \pi < 1, \\ (1 - \pi_i)\dfrac{\Gamma\left(y_i + \frac{1}{\tau}\right)(\tau\mu_i)^{y_i}}{y_i!\left(\frac{1}{\tau}\right)\ 1 + \tau\mu_i{}^{y_i} + \frac{1}{\tau}} & y_i > \tau,' \end{cases}$$

(9)

where μ_i is the mean of the NB model, π_i is the probability of a structural zero, τ Is the over-dispersion parameter, and Γ is the gamma function. The mean and variance of the ZINB are then given by:

$$E(y_i) = (1 - \pi_i)\mu_i$$

(10)

$$Var(y_i) = (1 - \pi_i)\mu_i(1 + \mu_i/\tau + \pi_i\mu_i)$$

(11)

As the parameter τ in the ZINB model approaches infinity, the model converges to the ZIP model. In contrast, smaller values of r indicate the presence of overdispersion in the data. Given this, is possible to say that ZINB model offers added flexibility compared to the ZIP model, by allowing for overdispersion arising from the excess zeros and heterogeneity in the Poisson component. On the other hand, the ZIP model only addresses overdispersion arising from excessive zeros, neglecting the potential heterogeneity in the Poisson component.

4.1.3. Hurdle Models

In contracts to zero-inflated models, hurdle models (Heilbron, 1994; Mullahy, 1986) consider zero values independently from count values and creates a different process, i.e., a different distribution, when generating zeros. This means that, contrary to Zero-inflated models, in Hurdle models only one process can produce zero count values. At the end of the estimation, the likelihood function is calculated by using the mixture of different distributions. Given this, Hurdle models can be viewed as a two-component mixture model consisting of a zero-mass modelled through logistic regression and the positive observations component following a truncated count distribution of the conventional models, such as truncated Poisson or truncated NB distribution.

Let Y_i denote the response of the ith observation, i = 1, …, n denote the total number of observations. The general structure of a hurdle model is given by:

$$P(Y_i = y_i) = \begin{cases} p_i & \text{if } y_i = 0, \\ (1 - p_i)\frac{p(y_i;\mu_i)}{1-p(y_i=0;\mu_i)} & \text{if } y_i > 0, \end{cases} \tag{12}$$

where p_i is the probability of a subject belonging to the zero component; $p(y_i;\mu_i)$ represents a probability mass function (PMF) for a regular count distribution with a vector of parameters μ_i and $p(y_i = 0;\mu_i)$ is the distribution evaluated at zero. It can be seen that the positive count is governed by a regular count distribution as the PMF divided by 1 minus the PMF of this regular count distribution evaluated at zero. In case the count distribution follows a Poisson distribution, the probability distribution for the *Hurdle Poisson (HP)* is given by:

$$P(Y_i = y_i) = \begin{cases} p_i & \text{if } y_i = 0, \\ (1 - p_i)\frac{e^{-\mu_i}\mu_i^{y_i}/y_i!}{1-e^{-\mu_i}} & \text{if } y_i > 0, \end{cases} \tag{13}$$

where μ_i is the mean of the Poisson model.

Alternatively, in case of existence of overdispersion not due to only excess zeros, the non-zero count component can follow a Negative Binomial (NB) distribution. The *Hurdle Negative Binomial (HNB)* is then given by:

$$P(Y_i = y_i) = \begin{cases} p_i & \text{if } y_i = 0, \\ \dfrac{1 - p_i}{1-\left(\frac{\tau}{\mu_i+\tau}\right)^{\tau}} \dfrac{\Gamma(y_i + r)}{\Gamma(\tau)y_i!}\left(\dfrac{\mu_i}{\mu_i+\tau}\right)^{y_i}\left(\dfrac{\tau}{\mu_i+\tau}\right)^{\tau} & \tau, \mu_i > 0; y_i > 0, \end{cases} \tag{14}$$

where $\mu_i$is the mean and $\mu_i\left(1+\frac{\mu_i}{\tau}\right)$ is the variance.

4.2. Model Evaluation and Performance Metrics

To assess the performance of the models, the dataset is partitioned into a training set and a test set. The training set is used to train the models, while the test set is employed to evaluate the models' accuracy on unseen data. To ensure a proper division of the dataset, the Train-test split technique is used to split the datasets. Common splits include 70/30, 80/20 or 90/10, where the first number represents the training set percentage, and the second number represents the test set percentage. Because our datasets are not very large, ideally it is desirable to represent as much as possible the different patterns on both train and test sets. For this reason, the method 70/30 is selected to reserve a larger proportion for testing and thus ensure a more reliable evaluation of the model's performance. Given that predictions will be done by year, a chronological ordering based on the year variable is applied and then, the 30% earliest records are assigned to the test set, while the remaining records are retained for the training set. This approach helps mitigate the issue of in-sample bias, where models tend to overfit the training data and subsequently perform poorly on new, unseen data (James et al., 2013).

For the evaluation of model performance on the training dataset, key metrics such as Log-Likelihood (LL) and Akaike Information Criterion (AIC) (Akaike, 2011) are utilized, which provide insights into the goodness-of-fit of the models. For the comparison of nested models Log-Likelihood Ratio Test (LRT) is used, followed by the Vuong Test (Vuong, 1989) to compare the non-nested models. Moreover, the evaluation extends to the test dataset, where additional metrics such as Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are employed to assess the accuracy of the models' predictions. These metrics provide valuable information on the predictive performance of the proposed models (Feng, 2021; Yildirim et al., 2022), complementing the analysis based on LL and AIC.

4.2.1. Log-Likelihood

The Log-Likelihood (LL) is a measure used in maximum likelihood estimation (MLE) to assess the fit of a statistical model to a determined dataset. In the MLE, the goal is to find the values of the model's parameters that maximize the likelihood of observing the given data. Hence, it quantifies the probability of observing the actual data under the assumed model, considering the estimated parameters. For this reason, the log-likelihood value provides a measure of how well the model fits the data. The log-likelihood is calculated by taking the natural logarithm of the likelihood function, which is proportional to the density of the data - the product of the probability densities of the observed data points (Agresti, 2019). Then the Log-Likelihood can be given by: $\ell(\theta; x) = \ln[L(\theta; x)]$.

In the scope of this study, the LL is used as a measure of the overall fit of the proposed models for the training data and hence to select the model that presents the best fit to the data. A higher log-likelihood value indicated a better fit, suggesting that the model captures the underlying patterns and characteristics of the data more accurately.

4.2.2. Akaike Information Criterion

The Akaike Information Criterion (AIC) is a measure of fit, which describes the performances of fitted models for a given dataset, and this will be used to compare the model fits between the conventional models, zero-inflated models, and hurdle models. Therefore, this measure is computed by $AIC = -2\log(L) + 2q$, where L is the likelihood, and q is the number of parameters in the model. In general, the best fitting model is the one that presents the lowest AIC values. An information criteria difference, which is less than 4 shows indifference between two models, between 4 and 10 differences indicate that one model is moderately superior to the other, and a difference greater than 10 suggests that one model is in reality better than the other (Chipeta et al., 2014).

4.2.3. Log Likelihood Ratio Test

The Log Likelihood Ratio Test (LRT) is used to compare nested models and it assesses whether adding or removing specific parameters significantly improves or reduces the fit of a model to the data (Agresti, 2019). More precisely, is a joint significance test. The test is based on comparing the log-likelihood values of two models: the full model and the reduced model. In the context of this study, the LRT is used to compare the conventional models, Poisson and NB, with the respective zero-inflated and hurdle models. To perform the Log Likelihood Ratio Test, first the parameters of the full model are estimated using maximum likelihood estimation (MLE). Followed by the parameters' estimation of the reduced model, which is a simplified version of the full model that excludes certain parameters. The log-likelihood is the logarithm of the likelihood function, which simplifies computations and allows for easier comparison of models. Therefore, the Log-Likelihood Ratio Test statistics equals to $2\log(\ell_1/\ell_0)$, where the ℓ_0 represents the likelihood of the reduced model, i.e., H_0 is true and ℓ_1 represents the likelihood for the full model, which is the MLE over all possible parameter values. If the alternative model provides a significantly better fit to the data than the null model, the LRT statistic will be larger and statistically significant, indicating a rejection of the null hypothesis in favor of the alternative hypothesis, i.e., the full model provides a significantly better fit to the data compared to the reduced model. Conversely, if the difference is small or not statistically significant, it indicates that the reduced model performs similarly to the full model and the additional parameters do not contribute significantly to the model's fit.

4.2.4. Vuong Test

The Vuong Test (Vuong, 1989) is a commonly used statistical tool for comparing non-nested models. It assesses the relative fit of two models by comparing their likelihood functions. In this context, let $f_1(y_i| x_1)$ and $f_2(y_i| x_2)$ represent the probability distribution functions of the two models. When both models fit the data equally well, their likelihood functions should be similar. Deviations between the likelihood functions indicate which model provides a better fit to the data.

To conduct the Vuong test, the maximum likelihood estimates (MLE) is obtained, denoted as $\widehat{x_1}$ and $\widehat{x_2}$, for the model parameters. The test then compares the likelihood functions at the MLEs for the two models. The Vuong test to compare $f_1(y_i| \widehat{x_1})$ and $f_2(y_i| \widehat{x_2})$ is then defined as $V = \sqrt{n}\bar{p}s_p$, where $\bar{p}$ and s_p is the mean and the standard deviation of the vector $p = (p_1, \dots, p_n)$. Under the null hypothesis, which states that the two models fit the data equally well, the Vuong statistic asymptotically follows a standard normal distribution. Using a significance level of 5%, the critical value is 1.96. So, if $V > 1.96$, the statistic favors the model in the numerator. Conversely, if $V < -1.96$, the statistic favours the model in the denominator, and when $V \in (-1.96, 1.96)$, it indicates that the two models fit the data equally well, without a preference for either model.

4.2.5. Root Mean Squared Error

The Root Mean Squared Error (RMSE) serves as an accuracy performance metric, calculated as the square root of the Mean Squared Error. The advantage of utilizing RMSE lies in its preservation of the original scale of the values, allowing for greater interpretability of the magnitude of the prediction error (Skiena, 2017). The Root Mean Squared Error is given by:

$$RMSE = \sqrt{\frac{\sum(y_i - \widehat{y_i})^2}{n}} \tag{15}$$

,where y_i is the actual values for the ith observation, $\widehat{y_i}$ is the predicted values for the ith observation and n is the number of observations.

4.2.6. MAE

The Mean Absolute Error (MAE) serves as a performance metric to evaluate the accuracy of predictions by quantifying the average absolute differences between the predicted values and the corresponding expected values. This metric offers the advantage of simplicity in interpretation and provides insight into the magnitude of the prediction error. It should be noted, however, that the MAE does not provide information regarding the direction of the error, as it is constrained to be always positive (Skiena, 2017). The Mean Absolute Error is then given by:

$$MAE = \frac{\sum_{i=1}^{n} | y_i - \widehat{y_i}|}{n}$$

(16)

,where y_i is the actual values for the th observation, $\widehat{y_i}$ is the predicted values for the ith observation, n is the number of observations.

5. RESULTS AND DISCUSSION

To investigate the performance of the regression models, each dataset has been split as train and test datasets, with the ratios of 70% and 30%, respectively. In addition to this, given that the datasets are aggregate on year level, before the split is done, the dataset is sorted by year, in order to make sure that the test dataset doesn't contain data from the "past". Hence the regression models were fitted using the train set and the test set to evaluate the performance of each model.

As mentioned previously, in the training stage five different models were fitted to each of the training set of the respective datasets: Poisson, Negative Binomial, Zero-Inflated Poisson, Zero-Inflated Negative Binomial, Hurdle Poisson and Hurdle Negative Binomial. To determine the model that best fits each training dataset, the metrics Akaike Information Criterion (AIC) and the Log Likelihood (LL) are used, and the model with higher LL and lower AIC is the one that represents the best goodness of fit for the datasets. Hence, in Table 12 the values of these metrics are presented for the respective training datasets, being possible to observe that overall, the models NB, ZINB and HNB are found to be the best models across the majority of the datasets. However, although the performances of these models are similar to each other, the best among them is the HNB model in the Consolidated, Goods & Supplies and Services datasets and the ZINB model in the Public Works dataset. Another point possible to observe is that the majority of the time the Negative Binomial models perform better than the Poisson models. Something to also notice is that the fit of the models is better for each contract type dataset, compared with the consolidated dataset that includes all the contracts.

Dataset / Models	Consolidated		Public Works		Goods & Supplies		Services	
	LL	AIC	LL	AIC	LL	AIC	LL	AIC
P	-25.608,84	51.243,68	-7.775,65	15.577,30	-10.849,75	21.725,51	-9.278,94	18.583,89
NB	-21.941,55	43.911,10	-6.777,52	13.583,04	-8.964,21	17.956,43	-8.276,32	16.580,63
ZIP	-23.747,93	47.547,87	-6.854,95	13.761,91	-10.099,62	20.251,25	-8.835,85	17.723,71
ZINB	-21.735,35	43.524,70	-6.523,06	13.100,11	-8.929,23	17.912,46	-8.220,93	16.495,85
HP	-23.762,06	47.576,11	-6.874,33	13.800,65	-10.100,30	20.252,61	-8.818,01	17.688,01
HNB	-21.686,53	43.427,05	-6.652,16	13.358,32	-8.827,54	17.709,08	-8.140,48	16.334,95

Table 12 - AIC and LL for Training Datasets

Therefore, to determine what model's variance can handle better the overdispersion in the data, a Log Likelihood Ratio (LRT) test has been carried out in the nested models. Thus, models based on negative binomial regression and poisson regression have been examined among themselves, separately. The results of this test are presented in Table 13, where is possible to observe that all comparisons of nested models have been statistically significant. This indicated that models based on negative binomial regression are more suitable when dealing with overdispersion and that these models have a significantly better fit than the poisson models.

Dataset / Models	Consolidated		Public Works		Goods & Supplies		Services	
	Value	p-value	Value	p-value	Value	p-value	Value	p-value
NB - P	7334,58	<0,001	1996,265	<0,001	3771,083	<0,001	2005,25	<0,001
ZINB - ZIP	4025,2	<0,001	663,8	<0,001	2340,8	<0,001	1229,9	<0,001
HNB - HP	4151,1	<0,001	444,3	<0,001	2545,5	<0,001	1355,1	<0,001

Table 13 - Log Likelihood Ratio Test on Training Datasets

Besides the LRT test, to compare non-nested models and determine which models presents a better performance when dealing with excessive zeros, a Vuong Test is applied in the training datasets. A positive value of test statistic in Vuong test indicates that the first model is more suitable than the second one. Table 14 presents the test statistics and significance values of the respective datasets. Hence, with Vuong test's results is possible to conclude that both Zero Inflated and Hurdle models are more suitable and statistically significant than the classical models Poisson and Negative Binomial. When comparing between zero-inflated and hurdle models, is possible to observe that in the Consolidated and Public Works datasets the zero-inflated models are preferable than the hurdle models and in the Goods & Supplies and Services datasets the Hurdle models are more suitable than the Zero Inflated models. The only case where this is not possible to prove is when comparing ZIP and HP for the Goods & Services dataset, given that the test is not statistically significant.

Dataset / Models	Consolidated		Public Works		Goods & Supplies		Services	
	z statistic	p-value	z statistic	p-value	z statistic	p-value	z statistic	p-value
P - ZIP	-18,47	< 2,22e-16	-15,60	< 2,22e-16	-10,91	< 2,22e-16	-9,10	< 2,22e-16
P - HP	-18,40	< 2,22e-16	-15,36	< 2,22e-16	-11,02	< 2,22e-16	-9,39	< 2,22e-16
ZIP - HP	1,09	0,14	1,74	0,09	0,17	0,43	-2,27	0,01
NB - ZINB	-9,22	< 2,22e-16	-11,45	< 2,22e-16	-3,45	<0,001	-4,19	<0,001
NB - HNB	-9,81	< 2,22e-16	-6,17	<0,001	-6,39	<0,001	-7,31	<0,001
ZINB - HNB	-1,60	0,06	6,54	<0,001	-4,38	<0,001	-3,94	<0,001

Table 14 - Vuong Test on Training Datasets

Following this, some performance criteria are also analyzed, but this time on unseen data, i.e., the testing datasets. Hence, the Root Mean Square Error (RMSE) and Mean Absolute Errors (MAE) are used in this analysis and presented in Table 15, for each test dataset. Based on the values shown in Table 15, is possible to see that although it represents the better results in the training dataset, HNB is usually the model that performs worse in the test dataset except for Public Works, given that NB is the models with worse results in this dataset. On the other side, the models that give the smaller RMSE and MAE are the ZIP and HP models, with very similar results in all datasets. In addition to this, is possible to observe that ZINB is the model that presents better results in the testing stage, from the all the models that presented better results in the training evaluation stage. Hence, although its prediction's accuracy is not so good as the poisson models, it still presents low deviation errors compared with NB and HNB.

Dataset / Models	Consolidated		Public Works		Goods & Supplies		Services	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
P	6,25	1,53	2,12	1,22	3,47	1,68	2,41	1,10
NB	958,86	11,95	5,73	1,52	4,92	1,76	2,44	1,12
ZIP	3,57	1,43	2,01	1,15	3,42	1,66	2,43	1,10
ZINB	118,45	2,78	2,08	1,15	4,93	1,76	2,65	1,13
HP	3,65	1,44	2,03	1,17	3,42	1,66	2,43	1,10
HNB	3.296,21	36,22	2,11	1,18	178,13	8,70	12,75	1,87

Table 15 - Performance Metrics in Testing Dataset

Put this, when the results from training and testing datasets are analyzed together, it's possible to observe that the model which performs better in the train phase are not the same as the models that perform better in the test phase. Usually, these cases suggest a trade-off between the model complexity and the predictive accuracy, given that the AIC and LL aims to strike a balance between the goodness of fit and the number of parameters in the model while the RMSE and MAE aims to measure the prediction error that assess the predictive performance of the models. Hence, a model with the best AIC, for example, achieves a good balance between the fit and complexity of the models and which allows it to capture the underlying patterns in the data and provide a reasonable representation of the relationships. On the other side, a model with the best RMSE and MAE means that it has a more refined understanding of the data and produces more precise predictions (Agresti, 2019). Given that one of the goals in this study is to understand the key factors that might influence the success of the companies in the public tenders, it can be said that the ZINB model is more preferable for inferential purposes and understanding of the relationships between the variables, but that at the same time has a lower prediction error compared with HNB, and thus the most suitable model for our analysis.

After selecting the appropriate model, the ZINB model was used to obtain the final fit and coefficients for each dataset. The modeling results for Public Works, Goods & Supplies, and Services can be found in Table 16 to Table 18, respectively. These tables consist of two parts: the Count part, which includes coefficients for companies that won one or more public tenders, and the Zero-inflated part, which includes coefficients for companies that have excess of zeros. The ZINB model captures both count and zero-dependent values using distinct distributions, allowing for different interpretations of the coefficients in each part. It's important to note that the coefficients in a ZINB model are typically reported in their logarithmic form. To interpret them in a more intuitive way, is possible to exponentiate the coefficients to obtain the corresponding incidence rate ratios for the count part, and the odds ratios for the zero-inflated part. Hence, in the count part an exponentiated coefficient greater than 1 indicates that the expected count of won tenders increases with a one-unit increase in the predictor variable, holding all other predictors constant, while an exponentiated coefficient less than 1 indicates a decrease in the expected count. For the zero-inflated component, an exponentiated coefficient greater than 1 suggests that the odds of excess zeros occurring, i.e., winning no public tenders, increase with a one-unit increase in the predictor variable, holding all other predictors constant, while an exponentiated coefficient less than 1 suggests a decrease in the odds of excess zeros. Therefore, by exponentiating the coefficients of the ZINB model, it's possible to interpret it as multiplicative effects on the odds or incidence rate, providing a more intuitive understanding of the relationship between the predictor variables and the excess zeros or count outcomes, offering a comprehensive understanding of the factors influencing the number of public tenders won. The exponential of the coefficients is also presented in Tables 16, 17 and 18 together with the Std Error and the Pr(>|z|). In this case, the Std Error represents the standard errors for the regression coefficients and dispersion parameter for the model, and the Pr(>|z|) represents the p-value and which defines the statistical significance of the respective independent variable. In addition to this, the log-likelihood and the likelihood-ratio chi-square test are also presented.

Variables	Count model coefficients					Zero-inflation model coefficients				
	β	exp(β)	Std Error	Pr(>|z|)		β	exp(β)	Std Error	Pr(>|z|)	
(Intercept)	-0,179	0,836	0,033	0,000	***	-3,563	0,028	0,488	0,000	***
Company_Size	0,194	1,215	0,036	0,000	***	0,188	1,206	0,108	0,082	.
Nbr_Employees	-0,099	0,906	0,037	0,007	**	-0,409	0,664	0,323	0,206	
ROE_PL_Before_Taxes_pct	0,032	1,033	0,037	0,392		-0,161	0,852	0,118	0,173	
Nbr_Current_Directors_Managers	-0,126	0,882	0,031	0,000	***	-0,308	0,735	0,116	0,008	**
ROCE_PL_Before_Taxes_pct	0,097	1,102	0,042	0,021	*	0,153	1,165	0,182	0,401	
ParticipatedTenders_Cum	0,436	1,546	0,021	< 2e-16	***	-8,069	0,000	0,924	< 2e-16	***
Company_Age	0,034	1,034	0,026	0,200		0,372	1,451	0,086	0,000	***
Current_Ratio	-0,124	0,883	0,037	0,001	***	-0,187	0,830	0,113	0,100	.
Solvency_Ratio_pct	0,021	1,022	0,030	0,470		-0,282	0,754	0,086	0,001	**
Total_Assets_EUR	0,049	1,050	0,045	0,279		0,360	1,433	0,225	0,110	
Operating_Revenue_EUR	-0,039	0,961	0,056	0,479		0,097	1,102	0,273	0,721	
Profit_Marging_pct	0,118	1,125	0,040	0,003	**	0,118	1,125	0,109	0,277	
Log(theta)	0,317	1,374	0,068	0,000	***	-	-	-	-	

Significance Codes:	0 '***'; 0.001 '**'; 0.01 '*'; 0.05 '.'; 0.1 ' '; 1
Nb. Of Observations:	6516 observations
Log-Likelihood:	-6506 on 27 Df
Likelihood-Ratio Test of Alpha=0:	Chisq = 2061,8; Pr(>Chisq) = < 2.2e-16

Table 16 - Fitting results of ZINB Model for Public Works Dataset

Variables	Count model coefficients					Zero-inflation model coefficients				
	β	exp(β)	Std Error	Pr(>|z|)		β	exp(β)	Std Error	Pr(>|z|)	
(Intercept)	-0,086	0,918	0,020	0,000	***	-8,754	0,000	1,665	<0,001	***
Company_Size	0,241	1,272	0,025	< 2e-16	***	1,440	4,219	0,408	0,000	***
Nbr_Employees	-0,116	0,890	0,022	0,000	***	1,435	4,200	0,386	0,000	***
ROE_PL_Before_Taxes_pct	0,056	1,058	0,024	0,019	*	0,418	1,518	0,196	0,033	*
Nbr_Current_Directors_Managers	-0,060	0,942	0,031	0,054	.	0,188	1,207	0,210	0,371	
ROCE_PL_Before_Taxes_pct	0,014	1,014	0,021	0,498		-0,046	0,955	0,164	0,778	
ParticipatedTenders_Cum	0,575	1,777	0,022	< 2e-16	***	0,609	1,838	0,146	0,000	***
Company_Age	0,071	1,073	0,020	0,000	***	-0,760	0,468	0,516	<0,001	
Current_Ratio	0,001	1,001	0,024	0,971		0,237	1,268	0,121	0,050	.
Solvency_Ratio_pct	-0,046	0,955	0,022	0,035	*	0,153	1,165	0,301	0,612	
Total_Assets_EUR	-0,016	0,984	0,027	0,564		1,791	5,996	0,641	0,005	**
Operating_Revenue_EUR	0,018	1,019	0,025	0,452		-28,064	0,000	7,094	0,000	***
Profit_Marging_pct	0,088	1,092	0,025	0,000	***	-0,034	0,967	0,336	0,920	
Log(theta)	-0,051	0,950	0,042	0,223		-	-	-	-	

Significance Codes:	0 '***'; 0.001 '**'; 0.01 '*'; 0.05 '.'; 0.1 ' '; 1
Nb. Of Observations:	6576 observations
Log-Likelihood:	-8791 on 27 Df
Likelihood-Ratio Test of Alpha=0:	Chisq = 1493,1; Pr(>Chisq) = < 2.2e-16 ***

Table 17- Fitting results of ZINB Model for Goods & Supplies Dataset

Variables	Count model coefficients					Zero-inflation model coefficients				
	β	exp(β)	Std Error	Pr(>|z|)		β	exp(β)	Std Error	Pr(>|z|)	
(Intercept)	-0,520	0,595	0,021	< 2e-16	***	-6,355	0,002	1,090	<0,001	***
Company_Size	0,234	1,264	0,026	< 2e-16	***	1,035	2,816	0,432	0,016	*
Nbr_Employees	0,003	1,003	0,018	0,845		-7,880	0,000	2,514	0,002	**
ROE_PL_Before_Taxes_pct	0,041	1,042	0,024	0,089	.	0,361	1,435	0,182	0,047	*
Nbr_Current_Directors_Managers	-0,023	0,977	0,021	0,275		-0,666	0,514	0,499	0,182	
ROCE_PL_Before_Taxes_pct	0,042	1,043	0,024	0,080	.	0,219	1,245	0,181	0,225	
ParticipatedTenders_Cum	0,418	1,519	0,022	< 2e-16	***	0,868	2,383	0,143	<0,001	***
Company_Age	0,015	1,016	0,019	0,422		-0,009	0,991	0,259	<0,001	
Current_Ratio	0,015	1,015	0,020	0,457		-0,525	0,591	0,861	0,542	
Solvency_Ratio_pct	0,026	1,027	0,022	0,231		0,592	1,808	0,382	0,121	
Total_Assets_EUR	0,032	1,032	0,020	0,106		0,810	2,248	0,488	0,097	.
Operating_Revenue_EUR	0,055	1,057	0,020	0,005	**	-3,384	0,034	2,274	0,137	
Profit_Marging_pct	0,023	1,024	0,024	0,327		-0,295	0,745	0,248	0,234	
Log(theta)	0,140	1,150	0,054	0,009	**	-	-	-	-	

Significance Codes:	0 '***'; 0.001 '**'; 0.01 '*'; 0.05 '.'; 0.1 ' '; 1
Nb. Of Observations:	7575 observations
Log-Likelihood:	-7990 on 27 Df
Likelihood-Ratio Test of Alpha=0:	Chisq = 1286,8; Pr(>Chisq) = < 2.2e-16 ***

Table 18- Fitting results of ZINB Model for Services Dataset

In Table 16, we present the results for the Public Works dataset, for both parts of the zero-inflated negative binomial (ZINB) model. This model presents a large likelihood-ratio test, and rejects the null hypothesis, suggesting that the predictors included in the full model have a statistically impact on the number of expected public tenders won. Focusing on the count part, the analysis reveals that several variables exhibit statistical significance in relation to winning public tenders. Notably, company size, number of current directors and managers, number of employees, return on capital employed (ROCE), cumulative number of participated tenders, current ratio and profit margin have significant effects. The results indicate that when the cumulative number of participated tenders by the company increases by one, the expected number of public work tenders won increases by a factor of 1.5. Additionally, the increase in one unit of the company size, i.e., from small to medium size for example, increases the expected count of won tenders by a factor of 1,2. Financial performance measures also play a crucial role in winning public tenders. Companies which profit margin or ROCE increases one percentual, have their expected number of won tenders increased by 1,1 factors. This suggests that companies that have effective cost management and efficient capital management have a greater chance of securing tenders. Conversely, when the current ratio increases one unit the expected number of won tenders decreases by 0,88 factor, which indicate that having a relatively higher current ratio compared to similar competitors may not reflect a good sign as, possibly indicating inefficient asset utilization. In the zero-inflated part of the analysis, several variables are found to influence the likelihood of a company winning zero public tenders or not. The statistically significant variables in this context include the number of current directors and managers, cumulative number of participated tenders, company age and solvency ratio. An increase of one tender in the cumulative number of participated tenders significantly decreases the odds of winning zero tenders in, highlighting the importance of past tendering experience in securing at least one tender, as was observed previously. Similarly, an increase of one unit in the number of directors and managers reduces the likelihood of winning zero tenders in 0,75 time, indicating that companies with a larger organizational structure have a greater chance of winning tenders. Interestingly, older companies have a higher likelihood of having excess zeros in their number of public tender wins, given that a company that is a year older has 1,5 higher odds of having zero public tenders won. This may suggest that older companies face challenges in winning public works tenders compared with younger companies, possibly due to factors such as innovation capacity or adaptation to changing procurement requirements. Lastly, financial measures such as solvency ratio demonstrate a positive impact on the performance of companies in winning public works tenders. An increase of one percentual in the solvency ratio decreases the likelihood of excess zeros by 0,75 times, suggesting that companies with a lower probability of default on long-term debts have a higher chance of winning tenders. These findings align with expectations, as financial stability and operational performance are crucial factors in the success of companies in public works tenders that usually represent bigger projects.

In Table 17, the coefficients for the count part and zero part of the zero-inflated negative binomial (ZINB) model applied to the Goods & Supplies tenders' dataset are presented, together with the respective log-likelihood and likelihood-ratio test. This model also presents a large likelihood-ratio test, and rejects the null hypothesis, suggesting that the predictors included in the full model have a statistically impact on the number of expected public tenders won. The count part reveals several statistically significant variables, including company size, number of employees, return on equity (ROE), cumulative number of participated tenders, company age, solvency ratio and profit margin. Notably, company size and company age have a positive impact, given that an increase of one unit in the size of the company, for example from small to medium, or an increase in one year on the age of the company increases the expected won tenders in goods and supplies contracts, by 1,3 and 1,1 times, respectively. This suggests that larger companies with greater resources, more experience and reputation in the market are more likely to win public tenders in this industry. These findings align with expectations, as bigger companies often possess higher production capacity and resources, providing them with a competitive advantage. This might also explain that companies that have an increase of one percent in Profit Margin or ROE percentages, have an expected increase of 1,1 times in the expected count of won tenders. This makes sense given that bigger companies can make use of economies of scale, and therefore be more cost efficient. Following this, an increase of one employee in the company will imply a decrease of 0,9 times in the expected awarded tenders. This might explain that companies whose production is more dependent on human labor rather than machinery, may have lower productivity and profitability. Moreover, in the zero-inflated part of the analysis, certain variables exhibit a dual effect on the outcome, whereby their increase is associated with both a higher number of public tenders won and an increased odd of excess zeros, implying a higher probability of not winning any public tenders. This pattern is observed for the variables company size, ROE and cumulative number of participated tenders, which suggests that these variables play a more complex role. On one side, they contribute to a higher expected count of won tenders, indicating a positive association with success. On the other hand, they also increase the probability of not winning any tenders, suggesting a potential source of zero inflation in the data. This could be due to factors such as competition, market conditions, or other unobserved variables that impact both the occurrence and frequency of the events of winning tenders. In particular the cumulative number of participated tenders has a very similar impact but in the opposite direction. In contrast, the number of employees aligns with the count part of the analysis, where the increase of this variable represents a significant increase in the odds of winning zero public tenders in this type of contract. Additionally, the operational revenue represents significant influence, where its increase indicates that that there is almost no chance of observing zero counts in the number of won public tenders. In contrast, the current ratio and total assets variables demonstrates a significant association with excess zeros in the dependent variable. An increase of these variables, independently, raises the odds of excess zeros, which might indicate that companies with a much higher current ratio associated with a big volume of assets are not using these efficiently.

Within the context of Services tenders, the coefficients for both the count and zero parts are examined in Table 18, together with the respective log-likelihood and likelihood-ratio test. This model also presents a large likelihood-ratio test, and rejects the null hypothesis, suggesting that the predictors included in the full model have a statistically impact on the number of expected public tenders won by the companies. Notably, the count part reveals a smaller number of statistically significant variables compared to other contract types. Company size, cumulative number of participated tenders and operational revenue are the statistically significant variables that positively influence the expected number of contracts awarded by companies in this sector. Among these, the cumulative number of participated tenders and company size have the most significant impact on the frequency of the dependent variable, where the increase of one unit in these variables represents increases the expected number of won tenders in the services contracts by 1,5 and 1,2 times, respectively. This might indicate that experience in public tender procedures and a larger organizational structure, which implies greater resources, are valued factors in the service industry. Also here, the operational revenue plays a positive role, given that an increase of one thousand euros in the Operational Revenue of a company increases the expected count of awarded tenders by 1 time. On the other hand, in the analysis of the zero-inflated coefficients, certain variables exhibit an opposite impact on the occurrence of excess zeros compared to the count part. Specifically, an increase in company size and cumulative number of tenders leads to a higher odd of encountering excess zeros in the number of won public tenders. Again, here the analysis of other factors and variables, like competition and market conditions, should be considered in order to understand the different outcomes in the dependent variable for these cases. On the other hand, the number of employees demonstrates a significant relationship with the dependent variable, as an increase of one employee in the staff of the company indicates that there is almost no chance of observing zero awarded tenders. Given that the services industry heavily relies on a company's resources, typically consisting of human labor, it is understandable that companies with a larger workforce would perform better in service tenders. Lastly, a higher solvency ratio and ROE can contribute to an excess of zeros in the dependent variable, given thar an increase of one percent in the solvency ratio or the ROE percentage increased the odds of observing zero counts of won tenders by 1.8 and 1.4 times, respectively. Although a high solvency ratio and ROE is generally indicative of financial stability, in the services market, it can also have downsides depending on the industry context. A high value for these metrics may imply a more conservative financial approach with reduced debts and limited leverage, which in turn restricts a company's ability to leverage resources for growth and innovation, particularly in technology-related industries, for example.

The analysis of the coefficients reveals a range of factors that impact the success of the Portuguese companies when biding in diverse types of public tenders, such as the company's experience, resource capacity, efficiency, profitability, and financial stability.

Hence, these findings can be compared to the existing studies reviewed in the literature review section. Despite there is a low number of studies that make use of business information to forecast the success of the companies in public tenders’ activity, is possible to identify similar patterns between our study’s results and the articles examined. The company size is one of the factors that demonstrated a higher impact, across all contract types, in the companies’ success in the public tenders, which is consistent with what was reviewed in other researches (Björkegren, 2022; Glas & Eßig, 2018). In addition to this, as also mentioned by these authors, our study also confirmed that despite bigger companies having higher success in public tenders, they usually also award the bigger projects, with longer duration and/or higher price. Furthermore, other factors outstanding in our analysis are also aligned with what we saw in further literatures, in particular the company’s experience in the market and in the public tenders’ procedures (Baranovsky et al., 2020), cost-effectiveness, resource capacity, innovation and financial stability (Bochenek, 2014; Lorentziadis, 2020; Cheaitou et al., 2019), are also presented as crucial non-price factors in the success of companies when participating in public tenders.

However, it is also important to note certain limitations in our study. Many previous analyses in the literature focus on variables related to tender information, which we could not include in our study due to the discrete nature of our dependent variable. Variables such as the number of competitors, geographical location, tender deadline, initial tender price, and number of entities can only be analyzed on a tender-specific basis, and therefore were not included in the current study. Nevertheless, our study offers a unique approach by merging a comprehensive set of business variables with a large number of contestants in public tenders in Portugal. Additionally, we break down the analysis by contract type to explore potential variations in impacting factors across different contract types. This distinguishes our study from previous research, which often focused on specific industries, companies, or a limited range of business data due to data availability challenges.

6. CONCLUSIONS

Public Procurement plays a relevant role in the economy, particularly within the private sector, as it aims to promote employment, innovation, sustainability, and the overall development of companies while ensuring efficient resource utilization. Hence, public procurement serves as a valuable tool for companies to participate in a competitive market, allowing for continuous improvement and development, to keep their survival in the market. Traditionally, the public procurement awarding process has been based on the lowest-price bid, as the primary criterion for awarding contracts. However, there has been a notable shift in recent years, with increasing awareness of the need to consider additional evaluation criteria beyond price alone.

Considering these factors, this study focuses on predicting the success of the Portuguese companies in the public tender procedures and identifying the key factors that influence their outcomes. Additionally, this study examines the factors of influence within each contract type to identify potential variations in these determinants across different types of contracts. The analysis provides valuable insights for companies intending to participate in such procedures, enabling them to assess their market position and develop more effective tendering strategies to increase their chances of winning public tenders. Moreover, this research holds benefits for public authorities, as it can inform the development of more objective evaluation criteria and contribute to a better understanding of the characteristics and criteria that lead to successful bids in the public tendering process.

To achieve the study goals, the public procurement data in Portugal is analyzed together with the business information of the participating companies in the tenders. The public procurement dataset includes historical information dating back to 2008, while the company dataset provides valuable insights into the characteristics of the tender contestants, including their experience, management, available resources, and financial stability. Through careful data extraction, cleaning, and merging processes, four separate datasets are created: consolidated, which includes all the public tenders; public works tenders; goods and supplies tenders and services tenders.

In the exploratory analysis stage, is possible to observe that the awareness of public tenders has significantly increased over the years, with a rise in the number of contestants and the total effective price awarded by companies. This surge in activity is accompanied by an increasing trajectory in operational revenue for winning companies, indicating a positive correlation between their performance and success in securing public tenders. In addition to this, it was possible to observe that the impact of both tender characteristics and contestant attributes varies across different contract types. Public Works tenders stand out due to their larger scale in terms of price volume, despite having comparatively shorter execution deadlines and higher levels of competition. In contrast, Goods/Supplies and Services contracts exhibit lower price values but longer durations and a less competitive environment. The statistics also shown that while small and medium-sized companies comprise the majority of contestants across all contract types, larger companies tend to participate in a greater number of tenders and secure contracts with higher monetary values and longer durations. Moreover, these larger companies exhibit higher success rates compared to their smaller counterparts. Lastly, an examination of business characteristics unveiled variables related to companies' experience, resource capacity, and financial performance that influence the number of tender wins. However, the specific impact of these factors varies depending on the contract type under consideration.

To measure the success of the companies in the public tender procedures, the number of public tenders won by each company per year is predicted. Given that the response variable is discrete in nature, count data regression methodologies were used. These models offer valuable insights and facilitate a comprehensive understanding of the relationship between the independent variables and the count outcome, thus enabling the identification of key factors influencing the companies' success in these procedures. Recognizing the existence of excess zero values in the response variable, advanced models such as Zero-Inflated and Hurdle models were also employed alongside the conventional Poisson and Negative Binomial models to better accommodate the excess of zeros and improve the accuracy of the predictions.

The analysis of model performance revealed that the more sophisticated models, namely the Zero-Inflated and Hurdle models, were better equipped to handle the overdispersion resulting from the excess of zeros in the data compared to the traditional Poisson and Negative Binomial models. Additionally, it was demonstrated that the Negative Binomial models outperformed the Poisson models in this dataset, effectively addressing the high variance that was not solely attributable to the excessive number of zeros. Consequently, the ZINB and HNB models emerged as the top-performing models in the training data, with the ZINB model demonstrating superior performance in the Public Works dataset and the HNB model exhibiting better performance in the Goods & Supplies and Services datasets. When evaluating the models using the test data, the ZIP and HP models demonstrated higher accuracy in predicting unseen data, while the HNB model performed less effectively. Considering the trade-off between goodness-of-fit and prediction accuracy, the ZINB model was considered to be the preferred choice for the subsequent inferential and coefficient analysis phase.

The coefficient analysis shows the various factors that influence the success of companies in different types of public tenders, including their experience, resource capacity, efficiency, profitability, and financial stability. Certain factors are common across all contract types, such as the company's experience in tender procedures and its size. Companies that have a long-term history of participating in public tenders and larger organizational structures are more likely to win a higher number of tenders. When comparing contract types, it becomes evident that public works contracts, characterized by larger projects, prioritize factors related to cost management, efficiency, and financial stability. These factors are particularly important due to the increased risks associated with these types of contracts for the contracting entities. In contrast, goods and supplies contracts, which usually have shorter durations, place less emphasis on metrics related to financial stability. Instead, factors such as company experience, size, reputation, and efficient resource utilization become more significant in these contracts, as they directly impact the company's production capacity. In services tenders, human resources appear as a crucial and positively impacting variable compared to other contract types. Additionally, market experience and innovative capabilities are identified as the most relevant factors for success in this specific tender category.

In conclusion, this study has successfully achieved its objectives by developing a model capable of addressing the excess of zeros and high variance in the datasets. The model exhibited a satisfactory level of accuracy, as evidenced by an average RMSE of 3.22 and MAE of 1.32. Furthermore, the study has identified and quantified the key factors that significantly influence the success of companies in the public tender procedure within each contract type. This comprehensive understanding of the determinants of success in public tenders provides valuable insights for companies, enabling them to tailor their strategies and improve their competitiveness in the market. Furthermore, this research holds significant benefits for public procurement authorities. By gaining insights into the factors that influence the success of companies in public tenders, authorities can elaborate more effectively the public procurement award criteria and respective weights in the award decision. Hence, this approach promotes a more competitive market environment where the selection process is not solely based on the lowest price bid and ensures that companies with diverse capabilities have the opportunity to secure public tenders, contributing to effective and equitable utilization of public resources. Moreover, policymakers can leverage these identified factors to develop targeted policies that support the growth and sustainability of companies facing market challenges, particularly small and medium-sized enterprises (SMEs). These efforts foster a competitive market environment that encourages innovation, economic development, and fair distribution of opportunities.

7. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

The current analysis showed several limitations that should be acknowledged. Firstly, the availability of public procurement data only provided information on the winning company, without accounting for multiple winners in a tender. As a result, there is a risk of misclassifying other contestants as non-winners, leading to an increased number of zeros in the data. Moreover, this data limitation prevents the identification and analysis of the impact of Lots within a tender, which has been explored in previous studies.

Another limitation is related to the lack of information on the announcement stage of public procurement, such as the time allocated for bid preparation, the range of lowest and highest bids, and the presence of co-operators. These variables have been examined in the reviewed literature and could potentially influence the success of companies. Their absence from the analysis restricts a comprehensive understanding of the factors at play during the tendering process.

For future research, it is suggested to employ a binary dependent variable approach to determine the probability of winning specific contracts based on the explanatory variables explored in this study, but also the variables related with the public tender characterization that had to be excluded in this case. This would involve utilizing machine learning models instead of count data regression models.

Additionally, incorporating business data related to stakeholders' information and municipal data could offer valuable insights. Social network analysis could be applied to examine how the relationships between a company's ultimate shareholders and the mayors of the public entities may influence their success in winning public procurement procedures. This approach would provide a deeper understanding of the dynamics and influences within the procurement ecosystem.

8. BIBLIOGRAPHY

Agresti, A. (2019). *Wiley Series in Probability and Statistics*. http://www.wiley.com/go/wsps

Akaike, H. (2011). Akaike's Information Criterion. In *International Encyclopedia of Statistical Science* (pp. 25–25). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-04898-2_110

Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015). A systematic review of open government data initiatives. *Government Information Quarterly*, *32*(4), 399–418. https://doi.org/10.1016/j.giq.2015.07.006

Ballesteros-Pérez, P., del Campo-Hitschfeld, M. L., Mora-Melià, D., & Domínguez, D. (2015). Modeling bidding competitiveness and position performance in multi-attribute construction auctions. *Operations Research Perspectives*, *2*, 24–35. https://doi.org/10.1016/j.orp.2015.02.001

Ballesteros-Pérez, P., González-Cruz, M. C., & Cañavate-Grimal, A. (2013). On competitive bidding: Scoring and position probability graphs. *International Journal of Project Management*, *31*(3), 434–448. https://doi.org/10.1016/j.ijproman.2012.09.012

Ballesteros-Pérez, P., Skitmore, M., Pellicer, E., & Gutiérrez-Bahamondes, J. H. (2016). Improving the estimation of probability of bidder participation in procurement auctions. *International Journal of Project Management*, *34*(2), 158–172. https://doi.org/10.1016/j.ijproman.2015.11.001

Ballesteros-Pérez, P., Skitmore, M., Sanz-Ablanedo, E., & Verhoeven, P. (2019). Forecasting the Number and Distribution of New Bidders for an Upcoming Construction Auction. *Journal of Construction Engineering and Management*, *145*(10). https://doi.org/10.1061/(asce)co.1943-7862.0001694

Baranovsky, A., Tkachenko, N., Glonti, V., Levchenko, V., Bogatyrova, K., Beridze, Z., Belinskaja, L., & Zelenitsa, I. (2020). Non-price criteria for the evaluation of the tender offers in public procurement of ukraine. *International Journal of Financial Studies*, *8*(3), 1–15. https://doi.org/10.3390/ijfs8030044

Bergman, M. A., & Lundberg, S. (2013). Tender evaluation and supplier selection methods in public procurement. *Journal of Purchasing and Supply Management*, *19*(2), 73–83. https://doi.org/10.1016/j.pursup.2013.02.003

Björkegren, E. (2022). *Predicting the size of a company winning a procurement: an evaluation study of three classification models*.

Bloomfield, K., Williams, T., Bovis, C., & Merali, Y. (2019). Systemic risk in major public contracts. *International Journal of Forecasting*, *35*(2), 667–676. https://doi.org/10.1016/j.ijforecast.2018.10.005

Bochenek, J. (2014). The contractor selection criteria in open and restricted procedures in public sector in selected EU countries. *Procedia Engineering*, *85*, 69–74. https://doi.org/10.1016/j.proeng.2014.10.530

Cameron, A. C. (Adrian C., & Trivedi, P. K. (1998). *Regression analysis of count data*. Cambridge University Press.

Chatfield, S. E. C., & Zidek, J. (2002). *Chapman & Hall/CRC - Texts in Statistical Science Series*.

Cheaitou, A., Larbi, R., & Al Housani, B. (2019). Decision making framework for tender evaluation and contractor selection in public organizations with risk considerations. *Socio-Economic Planning Sciences*, *68*. https://doi.org/10.1016/j.seps.2018.02.007

Cheung, Y. Bin. (2002). Zero-inflated models for regression analysis of count data: A study of growth and development. *Statistics in Medicine*, *21*(10), 1461–1469. https://doi.org/10.1002/sim.1088

Chipeta, M. G., Ngwira, B. M., Simoonga, C., & Kazembe, L. N. (2014). Zero adjusted models with applications to analysing helminths count data. In *BMC Research Notes* (Vol. 7). http://www.biomedcentral.com/1756-0500/7/856

Consul, P. C., & Famoye, F. (1992). Generalized poisson regression model. *Communications in Statistics - Theory and Methods*, *21*(1), 89–109. https://doi.org/10.1080/03610929208830766

Cox, D. R. (1983). Some remarks on overdispersion. In *Biometrika* (Vol. 70, Issue 1). http://biomet.oxfordjournals.org/

Curado, A., Damasio, B., Encarnacão, S., Candia, C., & Pinheiro, F. L. (2021). Scaling behavior of public procurement activity. *PLoS ONE*, *16*(12 December). https://doi.org/10.1371/journal.pone.0260806

Damásio, B., Louçã, F., & Nicolau, J. (2018). The changing economic regimes and expected time to recover of the peripheral countries under the euro: A nonparametric approach. *Physica A: Statistical Mechanics and Its Applications*, *507*, 524–533. https://doi.org/10.1016/j.physa.2018.05.089

Damásio, B., & Mendonça, S. (2019). Modelling insurgent-incumbent dynamics: Vector autoregressions, multivariate Markov chains, and the nature of technological competition. *Applied Economics Letters*, *26*(10), 843–849. https://doi.org/10.1080/13504851.2018.1502863

Damásio, B., & Nicolau, J. (2014). Combining a regression model with a multivariate Markov chain in a forecasting problem. *Statistics and Probability Letters*, *90*(1), 108–113. https://doi.org/10.1016/j.spl.2014.03.026

Damásio, B., & Nicolau, J. (2020). *Time Inhomogeneous Multivariate Markov Chains: Detecting and Testing Multiple Structural Breaks Occurring at Unknown*. https://www.facebook.com/researchrem/

Dean, C. B. (1992). Testing for overdispersion in poisson and binomial regression models. *Journal of the American Statistical Association*, *87*(418), 451–457. https://doi.org/10.1080/01621459.1992.10475225

Deepak K. Agarwal, Alan E. Gelfand, & Steven Citron-Pousty. (2002). Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics*, *9*, 341–355.

Dotoli, M., Epicoco, N., & Falagario, M. (2020). Multi-Criteria Decision Making techniques for the management of public procurement tenders: A case study. *Applied Soft Computing Journal*, *88*. https://doi.org/10.1016/j.asoc.2020.106064

European Commission. (2018). *Public Procurement Guidance For Practitioners*. https://doi.org/10.2776/461701

Falagario, M., Sciancalepore, F., Costantino, N., & Pietroforte, R. (2012). Using a DEA-cross efficiency approach in public procurement tenders. *European Journal of Operational Research*, *218*(2), 523–529. https://doi.org/10.1016/j.ejor.2011.10.031

Feng, C. X. (2021). A comparison of zero-inflated and hurdle models for modeling zero-inflated count data. In *Journal of Statistical Distributions and Applications* (Vol. 8, Issue 1). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1186/s40488-021-00121-4

Fry, T. D., Leitch, R. A., Philipoom, P. R., & Tian, Y. (2016). Empirical Analysis of Cost Estimation Accuracy in Procurement Auctions. *International Journal of Business and Management*, *11*(3), 1. https://doi.org/10.5539/ijbm.v11n3p1

García Rodríguez, M. J., Montequín, V. R., Fernández, F. O., & Villanueva Balsera, J. M. (2020). Bidders recommender for public procurement auctions using machine learning: Data analysis, algorithm, and case study with tenders from Spain. *Complexity*, *2020*. https://doi.org/10.1155/2020/8858258

García Rodríguez, M. J., Rodríguez Montequín, V., Ortega Fernández, F., & Villanueva Balsera, J. M. (2019). Public Procurement Announcements in Spain: Regulations, Data Analysis, and Award Price Estimator Using Machine Learning. *Complexity*, *2019*. https://doi.org/10.1155/2019/2360610

Glas, A. H., & Eßig, M. (2018). Factors that influence the success of small and medium-sized suppliers in public procurement: evidence from a centralized agency in Germany. *Supply Chain Management*, *23*(1), 65–78. https://doi.org/10.1108/SCM-09-2016-0334

Greenia, D. B., Qiao, M., & Akkiraju, R. (2014). A win prediction model for IT outsourcing bids. *Annual SRII Global Conference, SRII*, 39–42. https://doi.org/10.1109/SRII.2014.16

Grega, M., & Nemec, J. (2015). Factors Influencing Final Price of Public Procurement: Evidence from Slovakia. *Procedia Economics and Finance*, *25*, 543–551. https://doi.org/10.1016/s2212-5671(15)00768-6

Grzyl, B., Apollo, M., Miszewska-Urbańska, E., & Kristowski, A. (2018). The criteria for evaluation and selection the best tender applied by the authorities in Poland and selected EU countries. *MATEC Web of Conferences*, *219*. https://doi.org/10.1051/matecconf/201821904006

Hattori, T. (2010). Determinants of the number of bidders in the competitive procurement of electricity supply contracts in the Japanese public sector. *Energy Economics*, *32*(6), 1299–1305. https://doi.org/10.1016/j.eneco.2010.02.009

Heilbron, D. C. (1994). Zero-Altered and Other Regression Models for Count Data with Added Zeros. In *Biom. J* (Vol. 36).

Hilbe, J. M. (2011). *Negative Binomial Regression*. Cambridge University Press. https://doi.org/10.1017/CBO9780511973420

IMPIC. (2008). Código dos Contratos Públicos. *Decreto-Lei n.º 18/2008* .

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *Springer Texts in Statistics An Introduction to Statistical Learning*. http://www.springer.com/series/417

Jap, S. D., & Naik, P. A. (2008). BidAnalyzer: A method for estimation and selection of dynamic bidding models. *Marketing Science*, *27*(6), 949–960. https://doi.org/10.1287/mksc.1080.0363

Jones, S., & Wang, T. (2019). Predicting private company failure: A multi-class analysis. *Journal of International Financial Markets, Institutions and Money*, *61*, 161–188. https://doi.org/10.1016/j.intfin.2019.03.004

Lambert, D. (1992). Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing. In *Control TECHNOMETRICS* (Vol. 34, Issue 1).

Lenderink, B., Halman, J. I. M., & Voordijk, H. (2019). Innovation and public procurement: from fragmentation to synthesis on concepts, rationales and approaches. *Innovation: The European Journal of Social Science Research*, *35*(4), 650–674. https://doi.org/10.1080/13511610.2019.1700101

Lorentziadis, P. L. (2020). Competitive bidding in asymmetric multidimensional public procurement. *European Journal of Operational Research*, *282*(1), 211–220. https://doi.org/10.1016/j.ejor.2019.09.005

Lyra, M. S., Curado, A., Damásio, B., Bação, F., & Pinheiro, F. L. (2021). Characterization of the firm–firm public procurement co-bidding network from the State of Ceará (Brazil) municipalities. *Applied Network Science*, *6*(1). https://doi.org/10.1007/s41109-021-00418-y

Lyra, M. S., Damásio, B., Pinheiro, F. L., & Bacao, F. (2022). Fraud, corruption, and collusion in public procurement activities, a systematic literature review on data-driven methods. In *Applied Network Science* (Vol. 7, Issue 1). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s41109-022-00523-6

Maciej Malara, & Mariusz Mazurkiewicz. (2012). Modelling he Determinants Of Winning In Public Tendering Procedures Based On The Activity Of A Selected Company. *Operations Research and Decisions*, *22*, 51–62. https://doi.org/10.5277/ord120103

Marilena, M., & Alina, T. (2015). The Significance of Financial and Non-financial Information in Insolvency Risk Detection. *Procedia Economics and Finance*, *26*, 750–756. https://doi.org/10.1016/s2212-5671(15)00834-5

Martins, D., & Damásio, B. (2020). One Troika fits all? Job crash, pro-market structural reform and austerity-driven therapy in Portugal. *Empirica*, *47*(3), 495–521. https://doi.org/10.1007/s10663-019-09433-w

Mullahy, J. (1986). Specification and Testing of Some Modified Count Data Models. In *Journal of Econometrics* (Vol. 33).

Neelon, B., Ghosh, P., & Loebs, P. F. (2012). A spatial Poisson hurdle model for exploring geographic variation in emergency department visits. In *J. R. Statist. Soc. A* (Vol. 175). http://www.blackwellpublishing.com/rss

Neelon, B. H., O'Malley, A. J., & Normand, S. L. T. (2010). A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. *Statistical Modelling*, *10*(4), 421–439. https://doi.org/10.1177/1471082X0901000404

Neelon, B., O'Malley, A. J., & Smith, V. A. (2016). Modeling zero-modified count and semicontinuous data in health services research part 2: case studies. *Statistics in Medicine*, *35*(27), 5094–5112. https://doi.org/10.1002/sim.7063

Nicholas, C., & Fruhmann, M. (2014). Small and medium-sized enterprises policies in public procurement: Time for a rethink? *Journal of Public Procurement*, *14*(3), 328–360. https://doi.org/10.1108/jopp-14-03-2014-b002

Obwegeser, N., & Müller, S. D. (2018). Innovation and public procurement: Terminology, concepts, and applications. *Technovation*, *74–75*, 1–17. https://doi.org/10.1016/j.technovation.2018.02.015

OCDE. (2013). *Discussion Paper On Public Procurement Performance Measures*.

OECD. (2021). Size of public procurement. *Government at a Glance 2021, OECD Publishing, Paris*. https://doi.org/10.1787/888934258363

Ovsyannikova, A., & Domashova, J. (2020). Identification of public procurement contracts with a high risk of non-performance based on neural networks. *Procedia Computer Science*, *169*, 795–799. https://doi.org/10.1016/j.procs.2020.02.161

Pamučar, D., Božanić, D., Puška, A., & Marinković, D. (2022). APPLICATION OF NEURO-FUZZY SYSTEM FOR PREDICTING THE SUCCESS OF A COMPANY IN PUBLIC PROCUREMENT. *Decision Making: Applications in Management and Engineering*, *5*(1), 135–153. https://doi.org/10.31181/dmame0304042022p

Payne, E. H., Hardin, J. W., Egede, L. E., Ramakrishnan, V., Selassie, A., & Gebregziabher, M. (2017). Approaches for dealing with various sources of overdispersion in modeling count data: Scale adjustment versus modeling. *Statistical Methods in Medical Research*, *26*(4), 1802–1823. https://doi.org/10.1177/0962280215588569

Perumean-Chaney, S. E., Morgan, C., McDowall, D., & Aban, I. (2013). Zero-inflated and overdispersed: What's one to do? *Journal of Statistical Computation and Simulation*, *83*(9), 1671–1683. https://doi.org/10.1080/00949655.2012.668550

Rietveld, J., & Schilling, M. A. (2021). Platform Competition: A Systematic and Interdisciplinary Review of the Literature. *Journal of Management*, *47*(6), 1528–1563. https://doi.org/10.1177/0149206320969791

Rodriguez, M. J. G., Montequin, V. R., Ubierna, A. A., Hermida, R. S., Araujo, B. S., & Jauregi, A. Z. (2021). Award Price Estimator for Public Procurement Auctions Using Machine Learning Algorithms: Case Study with Tenders from Spain. *Studies in Informatics and Control*, *30*(4), 67–76. https://doi.org/10.24846/V30I4Y202106

Rose, C. E., Martin, S. W., Wannemuehler, K. A., & Plikaytis, B. D. (2006). On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *Journal of Biopharmaceutical Statistics*, *16*(4), 463–481. https://doi.org/10.1080/10543400600719384

Samprit Chatterjee, & Ali S. Hadi. (2015). Regression Analysis by Example. *John Wiley & Sons*.

Sarter, E. K., Fuchs, S., & Sack, D. (2014). *SME-friendly public procurement in Germany? Legal strategies and political challenges*. https://doi.org/10.13140/RG.2.1.4337.8165

Sedita, S. R., & Apa, R. (2015). The impact of inter-organizational relationships on contractors' success in winning public procurement projects: The case of the construction industry in the Veneto region. *International Journal of Project Management*, *33*(7), 1548–1562. https://doi.org/10.1016/j.ijproman.2015.05.001

SIGMA. (2016). Market Analysis, Preliminary Market Consultations, and Prior Involvement of Candidates/Tenderers. *Brief 32, OECD Publishing, Paris*. http://www.sigmaweb.org/publications/Public-Procurement- Policy-Brief-32-200117.pdf

Skiena, S. S. (2017). *The Data Science Design Manual*. Springer International Publishing. https://doi.org/10.1007/978-3-319-55444-0

Stiti, K., & Yape, S. J. (2019). *Bid Forecasting in Public Procurement*.

Técnica, F. (2021). *Relatório Anual 2021 Contratação Pública em Portugal 2021*. www.base.gov.pt

Telles, P. (2013). *The Good, the Bad and the Ugly: EU's Internal Market, Public Procurement Thresholds and Cross-Border Interest*.

Titl, V., Geys, B., Aisbett, E., Ashworth, S., Baltrunaite, A., Baranek, B., Bohn, F., Bracco, E., Breyer, F., Clinton, J. D., Colonnelli, E., Coviello, D., De Bruyne, G., D'inverno, G., Miceli, A. Di, De Witte, K., Farvaque, E., Hefeker, C., Horn, D., … Ursprung, H. (2018). *Political Donations and the Allocation of Public Procurement Contracts*.

Twizeyimana, J. D., & Andersson, A. (2019). The public value of E-Government – A literature review. In *Government Information Quarterly* (Vol. 36, Issue 2, pp. 167–178). Elsevier Ltd. https://doi.org/10.1016/j.giq.2019.01.001

Uyarra, E., & Flanagan, K. (2010). Understanding the innovation impacts of public procurement. *European Planning Studies*, *18*(1), 123–143. https://doi.org/10.1080/09654310903343567

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28. https://doi.org/10.1257/jep.28.2.3

Vaz, E., Bação, F., Damásio, B., Haynes, M., & Penfound, E. (2021). Machine learning for analysis of wealth in cities: A spatial-empirical examination of wealth in Toronto. *Habitat International*, *108*. https://doi.org/10.1016/j.habitatint.2021.102319

Vaz, E., Cusimano, M. D., Bação, F., Damásio, B., & Penfound, E. (2021). Open data and injuries in urban areas—A spatial analytical framework of Toronto using machine learning and spatial regressions. *PLOS ONE*, *16*(3), e0248285. https://doi.org/10.1371/journal.pone.0248285

Vaz, E., Damásio, B., Bação, F., Kotha, M., Penfound, E., & Rai, S. K. (2021). Mumbai's business landscape: A spatial analytical approach to urbanisation. *Heliyon*, *7*(7), e07522. https://doi.org/10.1016/j.heliyon.2021.e07522

Vuong, Q. H. (1989). Likelihood Ratio Test for Model Selection and Non-Nested Hypotheses. In *This content downloaded on Tue* (Vol. 57, Issue 2).

Yau, K. K. W., & Lee, A. H. (2001). Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. *Statistics in Medicine*, *20*(19), 2907–2920. https://doi.org/10.1002/sim.860

Yildirim, G., Kaciranlar, S., & Yildirim, H. (2022). Poisson and negative binomial regression models for zero-inflated data: an experimental study. *Communications Faculty Of Science University of Ankara Series A1Mathematics and Statistics*, *71*(2), 601–615. https://doi.org/10.31801/cfsuasmas.988880

Yu, Q., Miche, Y., Séverin, E., & Lendasse, A. (2014). Bankruptcy prediction using Extreme Learning Machine and financial expertise. *Neurocomputing*, *128*, 296–302. https://doi.org/10.1016/j.neucom.2013.01.063

APPENDIX

Appendix A

Name	Description
Tender ID	Identifier of the contract
Title	Brief description of the contract
Initial Contractual Price	Initial awarded price of the contract
Total Effective Price	Total effective price at the end of the contract
Publication Date	Date of the contract announcement
Signing Date	Date of award contract signature
Close Date	Award contract end date
Contracted NIF	Identifier of the winning bidder: NIF (Tax Number)
Contracting	Identifier of the contracting entity: NIF (Tax Number)
Announcement ID	Identifier of the original contract's announcement publication
Direct Award Fundamentation Type	PPC Article that fundaments the choice of direct ward contracting procedure
Invitees	Identifier of companies invited by the contracting entity in case of a direct agreement or restricted tender
Observations	Further observations regarding the contract
Type of Contract End	Description of the contract end type
Contract Fundamentation Type	PPC Article that fundaments the contract type choice by the contratcing entity
Causes of Deadline Changes	Description of reasons why the contract's deadline was changed
Framework Agreement Procedure Description	Description of the framework agreement utilized in the procedure
Contracting Procedure Type	Procedure by which the contract was announced: direct award under the general regime, public tender, restricted tender with prior qualification, negotiation procedure, prior consultation, restricted tender with prior qualification, competitive dialogue, innovation partnership
Contract Type	Type of contract defined by legislation (Article 16th of PPC): public works contract; public works concession; public services concession; leasing or acquisition of movable assets; acquisition of services; society.
Execution Deadline	Number of days to conclude the work agreed in the contract
Execution Place	Contract's execution place: Country, District and City
Centralized Procedure	Indicates weather it's a centralized procedure or not
Description	Lond description of the contract's object and context
Contract Status	Status of the contract
Contestant NIF	Identifier of the companies that bid in the tender: NIF (Tax number)
Nbr of Contestans	Number of total companies that participated in the procurement.
Creation Date	Date in which the contract was created in BASE database
Last Update Date	Last date in which the contract was updated
Count of Updates	Number of times the contract information was updated
Price Changes Causes	Description of causes for the price change
Co-contratantes	Indicates whether there is the presence of co-contratants or not
CPV	Numerical code of 8 digits that identifies the nature of the work and type of contract

Table 19 - Public Procurement Fields Description

Appendix B

Name	Description
Company name	Name of the company
Company Tax Number	The company's tax number provides formal registration on the company tax system in Portugal (NIF)
Date of incorporation	Is the date on which the company starts its activity
Number of directors & managers	Total number of directors and managers that existed in the company
Number of current directors & managers	Total number of current directors and managers that exist in the company for the running year
Number of previous directors & managers	Total number of directors and managers that exist in the company in the past
Directing Manager Name	Name of the managers of the company
Directing Manager Gender	Gender of the manager of the company
Directing Manager Birth date	Birth Date of the manager of the company
Directing Manager Job Title	Job Title of the manager of the company
Directing Manager is Also a Shareholder	Indicates whether the manager is also a stakeholder (Yes or No)
Number of Shareholders	Number of stakeholders of the company
Shareholder Name	Name of stakeholders of the company
Shareholder - Total %	This corresponds to the percentage of indirect ownership. In this type of structure, a simple chain of two or more ownership links exists between an owned or part-owned entity, and an owning or part-owning individual or entity, via one or more intermediate entities
Shareholder - Direct %	This corresponds to the percentage of direct ownership. An entity is directly owned or part-owned by an individual or entity if the ownership link is direct and passes through no other individuals or entities
Shareholder Type	Indicates whether the shareholder corresponds do an individual or corporation
Shareholder is Also a manager	Indicates whether the shareholder is also a manager of the company (Yes or No)
NACE Rev. 2	European industrial activity classification
Year	Year to each the information is related to
Closing date	Corresponds to the Accounting Closing Date
Fiscal year Last avail. yr	Last fiscal year information available for the company
Audit status	The audit status indicates the status of a statement with regard to its auditing: Qualified, Unqualified, Unaudited, Adverse opinion, No opinion, Qualification n.a.
Operating revenue (Turnover) th EUR	Total operating revenues (Net sales + Other operating revenues + Stock variations) in thousand euros. The figures do not include VAT or excises taxes and similar obligatory payments.
P/L before tax th EUR	Equals the total of operating profits and financial profits in thousand euros
P/L for period [=Net income] th EUR	Net income for the Year in thousand euros. Before deduction of Minority interests if any (profit after taxation + extraordinary and toher profit)
Cash flow th EUR	Equals to the profit of the period plus depreciation in thousand euros.
Total assets th EUR	Equals to the fixed assets plus current assets in thousand euros
Shareholders funds th EUR	Total equity, that equals to the Capital plus Other shareholders funds in thousand euros.
Current ratio	Equals to Current Assets / Current Liabilities
Profit margin	(Profit Before Taxes / Operating Revenue) *
ROE using P/L before tax %	(Profit before tax / Shareholders funds) * 100
ROCE using P/L before tax %	(Profit before tax + Interest paid) / (Shareholders funds + Non-current liabilities) * 100
Solvency ratio (Asset based) %	(Shareholders funds / Total assets) * 100
Number of employees	Total number of full-time employees of the company (personnel)

Table 20 - Companies Dataset Fields Description

Appendix C

Variable	Missing Values %
Tender_ID	0,00
initial_contractual_price	0,00
close_date	75,87
contracting	0,00
invitees	0,00
publication_date	0,00
end_of_contract_type	73,42
total_effective_price	0,00
contracting_procedure_type	0,00
contract_types	0,00
execution_deadline	0,00
execution_place	0,00
contract_status	100,00
signing_date	0,00
created_at	0,00
last_updated_at	0,00
updated_count	0,00
cocontratantes	100,00
cpvs	0,00
Contracted_NIF	0,00
NbrContestans	0,00
Contestant_NIF	0,00

Table 21 - Missing Values in Public Procurements Dataset

Appendix D

Variable	Missing Values %
VAT Number	0,00
Company_Name	0,00
Incorporation_Date	0,06
NACE_Rev2	0,10
LastAvail_FiscalYear	0,63
Year	0,00
Operating_Revenue_EUR	0,00
PL_Before_Taxes_EUR	0,00
PL_After_Taxes_EUR	0,00
Cash_Flow_EUR	0,02
Total_Assets_EUR	0,00
Shareholders_Funds_EUR	0,00
Current_Ratio	0,05
Profit_Margin_pct	0,37
ROE_PL_Before_Taxes_pct	0,15
ROCE_PL_Before_Taxes_pct	1,04
Solvency_Ratio_pct	0,06
Nbr_Employees	0,09

Table 22 - Missing Values in Companies Dataset

Appendix E

Field	Calculation
Won	IF Contestant_NIF = Contracted_NIF THEN 1 ELSE 0
NrOfContractingEntities	Count list elements in each row of 'contracting' columns
Year	YEAR(signing_date)
execution_deadline	IF Won = 1 THEN execution_deadline ELSE 0
tender_closed	IF Won = 1 THEN IF close_Date IS NOT NULL OR end_of_contract_type IS NOT NULL THEN 1 ELSE 0 ELSE 0
initial_contractual_price	IF Won = 1 THEN initial_contractual_price / 100 ELSE 0
total_effective_price	IF Won = 1 AND tender_closed = 1 THEN IF total_effective_price = -1 THEN initial_contractual_price / 1000 ELSE total_effective_price / 1000 ELSE 0
ContractualPrice_Deviation	IF tender_closed = 1 THEN total_effective_price - initial_contractual_price ELSE None
ContractualPrice_Deviation_pct	IF tender_closed = 1 THEN ContractualPrice_Deviation / initial_contractual_price * 100 ELSE None
Effective_ExecutionDays	IF tender_closed = 1 THEN IF close_date IS NULL THEN execution_deadline ELSE close_date - signing_date ELSE None
ExecutionDays_Deviation_pct	IF tender_closed = 1 AND execution_deadline IS NOT NULL AND execution_deadline <> 0 THEN ExecutionDays_Deviation / execution_deadline*100 ELSE None
contract_type	IF contract_types CONTAINS ("obras", "empreitadas") THEN 'Public Works' ELSE IF contract_types CONTAINS ("bens") THEN 'Goods/Supplies' ELSE IF contract_types CONTAINS ("serviços", "sociedade") THEN 'Services'' ELSE None
Nbr_ParticipatedTenders	Equals 1 for every Contestant that participated in the tender
Cumulative_ParticipatedTenders	Cumulative sum of the field Nbr_ParticipatedTenders since 2012 until the Year-1
Cumulative_ClosedTenders	Cumulative sum of the field tender_closed since 2012 until the Year-1
Cumulative_WonTenders	Cumulative sum of the field Wons since 2012 until the Year-1

Table 23 - New Variables Created in Public Procurements Dataset

Appendix F

Field	Calculation
Operating_Revenue/Nbr_Employees	Operating_Revenue * 1000 / Nbr_Employees
Company_Size	IF Operating Revenue EUR >= 100 million OR Total Assets EUR >= 200 million OR Nb. Employees >= 1000 AND Operating Revenue / Nb. Employees >= 100 THEN 'VL' ELSE IF Operating Revenue EUR >= 10 million OR Total Assets EUR >= 20 million OR Nb. Employees >= 150 AND Operating Revenue / Nb. Employees >= 100 THEN 'L' ELSE IF Operating Revenue EUR >= 1 million OR Total Assets EUR >= 2 million OR Nb. Employees >= 15 AND Operating Revenue / Nb. Employees >= 100 THEN 'M' ELSE 'S'
Cash_ROA	IF Total_Assets_EUR = 0 THEN 0 ELSE Cash_Flow_EUR/Total_Assets_EUR
ROA_PL_After_Taxes_pct	IF Total_Assets_EUR = 0 THEN 0 ELSE ROA_PL_After_Taxes_pct/Total_Assets_EUR
Incorporation_Year	YEAR(Incorporation_Date)
Company_Age	FiscalYear - Incorporation_Year

Table 24 - New Variables Created in Companies Dataset

Appendix G

Variables	count	mean	std	min	0,25	0,5	0,75	max
Year	27763	2016,03	2,29	2012	2014	2016	2018	2019
Won_Tenders	27763	1,31	3,31	0	0	0	1	134
Initial_Contractual_Price	27763	334,86	1858,55	0	0	0	139,03	82472,43
Total_Effective_Price	27763	333,06	1854,21	0	0	0	136,69	82472,43
Tenders_Closed	27763	0,39	1,14	0	0	0	0	28
Nbr_ParticipatedTenders	27763	8,94	20,27	1	1	3	8	723
ParticipatedTenders_Cum	27763	27,25	61,34	1	2	7	25	1649
Won_Cum	27763	3,87	10,45	0	0	1	3	410
TendersClosed_Cum	27763	1,47	4,13	0	0	0	1	107
Avg_Execution_Deadline	13261	368,3	1433,63	1	120	262	405	148432
Avg_NbrContestans	27763	9,77	6,9	1	5	8	13	66
Avg_NrOfContractingEntities	13261	1	0,17	1	1	1	1	17
Avg_ContractualPrice_Deviation_pct	5929	-4,29	20,14	-98	-4	0	0	241
Avg_ExecutionDays_Deviation_pct	5929	119,43	902,03	-99,66	-1,02	7,78	107,64	54650
Company_Age	27763	21,05	16,61	0	9	18	28	204
Company_Size	27763	1,87	0,79	1	1	2	2	4
Nbr_Current_Directors_Managers	27763	4,25	4,54	0	2	3	5	167
Nbr_Shareholders	27763	3,23	3,46	0	2	3	4	108
FiscalYear	27763	2015,03	2,29	2011	2013	2015	2017	2018
Operating_Revenue_EUR	27763	19840,16	199084,74	0	425	1349	4665,5	10848666
PL_Before_Taxes_EUR	27763	373,95	29110,18	-2651824	3	26	144	847847
PL_After_Taxes_EUR	27763	185,64	26479,55	-2271394	1	17	103	733210
Cash_Flow_EUR	27763	1273,65	22029,52	-1480078	9	51	216	965119
Total_Assets_EUR	27763	87953,1	2127830,32	0	407	1328	4550	112962840
Shareholders_Funds_EUR	27763	11225,33	178965,6	-339474	99	434	1625,5	8285445
Current_Ratio	27763	2,77	5,34	0	1,17	1,61	2,57	99,89
Profit_Marging_pct	27763	2,63	14,97	-99,82	0,63	2,42	6,67	100
ROE_PL_Before_Taxes_pct	27763	9,02	69,28	-989,28	2	8,65	22,29	965,56
ROCE_PL_Before_Taxes_pct	27763	7,79	56,24	-989,19	2,5	7,65	17,15	993,74
Solvency_Ratio_pct	27763	37,33	27,76	-99,87	20,14	35,31	55,72	100
Nbr_Employees	27763	109,92	636,02	1	6	15	42	19535
Revenue_Per_Employee	27763	527409,41	9839320,64	0	44500	79750	162408,5	643122000
Cash_ROA	27763	0,06	2,46	-31,75	0,02	0,05	0,1	346
ROA_PL_After_Taxes_pct	27763	0,02	2,23	-31,75	0	0,02	0,06	325,5

Table 25 - Consolidated Dataset Descriptive Statistics

Appendix H

Variables	count	mean	std	min	0,25	0,5	0,75	max
Year	9216	2015,71	2,31	2012	2014	2016	2018	2019
Won_Tenders	9216	0,95	1,92	0	0	0	1	25
Initial_Contractual_Price	9216	483,25	2039,75	0	0	0	261,48	82063,7
Total_Effective_Price	9216	482,99	2042,15	0	0	0	260,23	82063,7
Tenders_Closed	9216	0,3	0,85	0	0	0	0	12
Nbr_ParticipatedTenders	9216	8,02	13,06	1	1	3	9	169
ParticipatedTenders_Cum	9216	36,38	65,31	1	4	13	40	1258
Won_Cum	9216	4,25	10,43	0	0	1	4	410
TendersClosed_Cum	9216	1,72	4,08	0	0	0	2	76
Avg_Execution_Deadline	3586	202,37	222,61	8	101	162,5	258	9125
Avg_NbrContestans	9216	12,38	6,36	1	8	11	16	49
Avg_NrOfContractingEntities	3586	1	0	1	1	1	1	1
Avg_ContractualPrice_Deviation_pct	1684	-0,48	7,16	-90	-1,67	0	0	43
Avg_ExecutionDays_Deviation_pct	1684	153,18	233,38	-61,1	0	93,33	214,03	5232,5
Company_Age	9216	21,37	14,28	0	11	19	29	97
Company_Size	9216	1,86	0,68	1	1	2	2	4
Nbr_Current_Directors_Managers	9216	3,87	3,22	0	2	3	5	30
Nbr_Shareholders	9216	3,22	2,13	0	2	3	4	20
FiscalYear	9216	2014,71	2,31	2011	2013	2015	2017	2018
Operating_Revenue_EUR	9216	10011,4	83351,46	0	542,75	1423	4018,5	3277540
PL_Before_Taxes_EUR	9216	157,88	4876,93	-156770	4	22	104,25	127173
PL_After_Taxes_EUR	9216	72,88	4549,46	-173587	1	14	75	126425
Cash_Flow_EUR	9216	352,48	6685	-155682	11	50	177	464922
Total_Assets_EUR	9216	14494,92	107236,05	0	573,75	1598	4471,25	6870588
Shareholders_Funds_EUR	9216	4192,76	23133,74	-110009	177	538,5	1685	583404
Current_Ratio	9216	2,77	4,6	0	1,28	1,72	2,67	95,82
Profit_Marging_pct	9216	1,31	14,33	-99,82	0,57	1,91	5,16	100
ROE_PL_Before_Taxes_pct	9216	3,34	64,67	-989,28	1,36	5,66	16,12	908,62
ROCE_PL_Before_Taxes_pct	9216	4,72	49,86	-989,19	2,12	6,32	13,86	780,67
Solvency_Ratio_pct	9216	38,6	24,63	-99,87	22,53	36,12	54,78	100
Nbr_Employees	9216	63,83	283,4	1	9	20	43	7817
Revenue_Per_Employee	9216	275622,44	7811323,3	0	45000	71646,5	120431,75	556809000
Cash_ROA	9216	0,04	0,29	-8	0,01	0,04	0,08	16,5
ROA_PL_After_Taxes_pct	9216	0	0,28	-8	0	0,01	0,04	16,5

Table 26 - Public Works Dataset Descriptive Statistics

Appendix I

Variables	count	mean	std	min	0,25	0,5	0,75	max
Year	10755	2016,28	2,2	2012	2015	2017	2018	2019
Won_Tenders	10755	1,51	3,62	0	0	1	1	89
Initial_Contractual_Price	10755	173,99	1129,18	0	0	2,42	91,84	49859,08
Total_Effective_Price	10755	171,82	1125,99	0	0	2	89,92	49925,71
Tenders_Closed	10755	0,5	1,39	0	0	0	1	28
Nbr_ParticipatedTenders	10755	10,37	26,22	1	1	3	9	721
ParticipatedTenders_Cum	10755	39,13	83,08	1	3	11	36	1649
Won_Cum	10755	6,08	14,96	0	0	2	6	410
TendersClosed_Cum	10755	2,32	5,86	0	0	0	2	107
Avg_Execution_Deadline	5587	323,72	1891,56	1	60	221	365	139000
Avg_NbrContestans	10755	8,03	5,72	1	4	6	11	45
Avg_NrOfContractingEntities	5587	1	0,12	1	1	1	1	5
Avg_ContractualPrice_Deviation_pct	2774	-6,66	19,46	-97	-6,75	0	0	200
Avg_ExecutionDays_Deviation_pct	2774	124,17	1239,81	-99,59	-12,29	1,66	60,43	54650
Company_Age	10755	23,51	18,06	0	11	20	31	176
Company_Size	10755	2,01	0,81	1	1	2	2	4
Nbr_Current_Directors_Managers	10755	4,69	5,08	0	2	3	6	167
Nbr_Shareholders	10755	3,31	3,57	0	2	3	4	108
FiscalYear	10755	2015,28	2,2	2011	2014	2016	2017	2018
Operating_Revenue_EUR	10755	35401,34	305035,31	0	734,5	2250	7652,5	10848666
PL_Before_Taxes_EUR	10755	709,38	31318,2	-2360387	6	47	243	759278
PL_After_Taxes_EUR	10755	378,01	30765,8	-2271394	3	31	177	733210
Cash_Flow_EUR	10755	2067,32	27341,78	-1480078	17	81	337,5	965119
Total_Assets_EUR	10755	83875,06	1773441,24	0	619	2064	7214,5	100901467
Shareholders_Funds_EUR	10755	13720,21	173763,54	-174867	152	684	2637	6183709
Current_Ratio	10755	2,48	4,02	0	1,17	1,59	2,5	89,77
Profit_Marging_pct	10755	2,94	11,8	-99,82	0,72	2,52	6,25	99,02
ROE_PL_Before_Taxes_pct	10755	10,65	63,88	-976,55	2,67	9,57	22,29	913,97
ROCE_PL_Before_Taxes_pct	10755	9,68	52,67	-957,17	3,2	8,45	17,59	788,28
Solvency_Ratio_pct	10755	37,64	26,8	-98,21	20,23	35,21	56,06	100
Nbr_Employees	10755	103,22	553,63	1	6	15	44	18873
Revenue_Per_Employee	10755	885474,15	13537308,88	0	74417	137222	262287	556809000
Cash_ROA	10755	0,09	3,88	-30	0,02	0,05	0,09	346
ROA_PL_After_Taxes_pct	10755	0,06	3,51	-30	0	0,02	0,06	325,5

Table 27 - Goods and Supplies Dataset Descriptive Statistics

Appendix J

Variables	count	mean	std	min	0,25	0,5	0,75	max
Year	12314	2016,16	2,29	2012	2014	2017	2018	2019
Won_Tenders	12314	0,93	2,69	0	0	0	1	106
Initial_Contractual_Price	12314	241,34	1668,17	0	0	0	90,42	61599,14
Total_Effective_Price	12314	239,37	1660,27	0	0	0	89,9	61599,14
Tenders_Closed	12314	0,21	0,7	0	0	0	0	15
Nbr_ParticipatedTenders	12314	5,1	11,33	1	1	2	4	272
ParticipatedTenders_Cum	12314	25,69	62,94	1	2	6	21	1649
Won_Cum	12314	4,34	12,9	0	0	1	4	410
TendersClosed_Cum	12314	1,5	4,57	0	0	0	1	107
Avg_Execution_Deadline	5283	561,84	2146,33	1	223	365	730	148432
Avg_NbrContestans	12314	8,56	7,16	1	4	7	11	66
Avg_NrOfContractingEntities	5283	1,01	0,26	1	1	1	1	17
Avg_ContractualPrice_Deviation_pct	1729	-3,96	28,88	-98	-8,5	0	0	241
Avg_ExecutionDays_Deviation_pct	1729	72,13	529,55	-99,66	-3	3,2	39,04	18300
Company_Age	12314	19,82	17,17	0	9	17	26	204
Company_Size	12314	1,88	0,87	1	1	2	2	4
Nbr_Current_Directors_Managers	12314	4,64	4,98	0	2	3	6	60
Nbr_Shareholders	12314	3,26	4,19	0	2	2	4	85
FiscalYear	12314	2015,16	2,29	2011	2013	2016	2017	2018
Operating_Revenue_EUR	12314	30266,76	289812,93	0	288	1121	4899	10848666
PL_Before_Taxes_EUR	12314	424,56	43424,95	-2651824	3	25	160	847847
PL_After_Taxes_EUR	12314	142,24	39489,79	-2271394	1	16	114	733210
Cash_Flow_EUR	12314	2245,72	32626,73	-1480078	7	49	244	965119
Total_Assets_EUR	12314	183657,76	3191754,21	0	269	1078	4557	112962840
Shareholders_Funds_EUR	12314	20220,08	267412,45	-339474	62	326	1496,5	8285445
Current_Ratio	12314	2,84	6,24	0	1,1	1,52	2,45	99,89
Profit_Marging_pct	12314	3,54	16,63	-99,56	0,7	2,86	8,2	100
ROE_PL_Before_Taxes_pct	12314	11,65	74,11	-976,55	2,39	10,59	27,19	965,56
ROCE_PL_Before_Taxes_pct	12314	8,77	60,65	-988,23	2,55	8,46	19,16	993,74
Solvency_Ratio_pct	12314	36,14	29,35	-99,76	18,77	34,56	55,36	100
Nbr_Employees	12314	187,55	915,11	1	5	16	54	19535
Revenue_Per_Employee	12314	806422,49	14049132,56	0	34500	61522,5	112000	643122000
Cash_ROA	12314	0,05	0,66	-31,75	0,02	0,06	0,12	32,5
ROA_PL_After_Taxes_pct	12314	0	0,64	-31,75	0	0,02	0,07	32,25

Table 28 - Services Dataset Descriptive Analysis

Appendix K

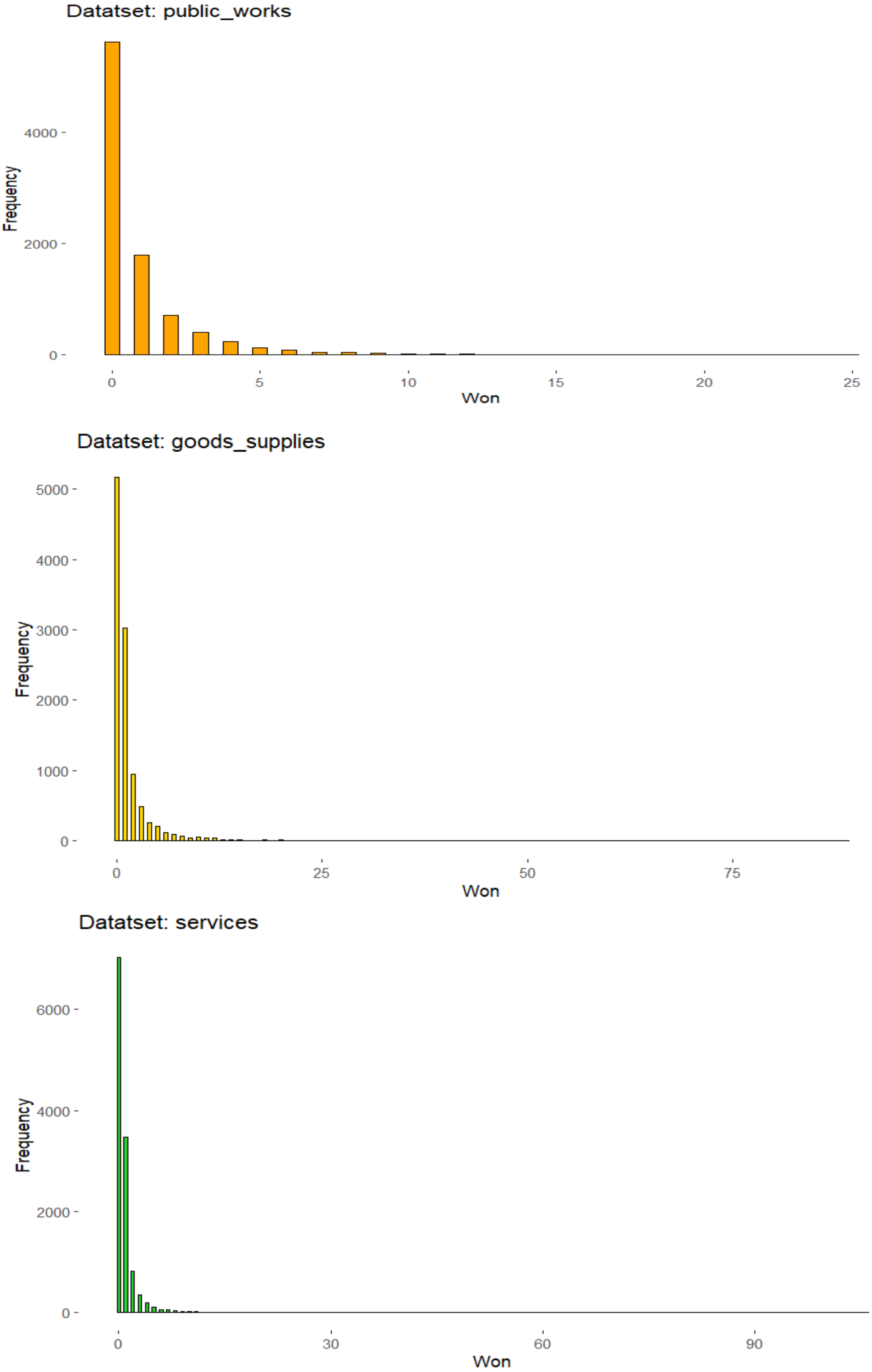
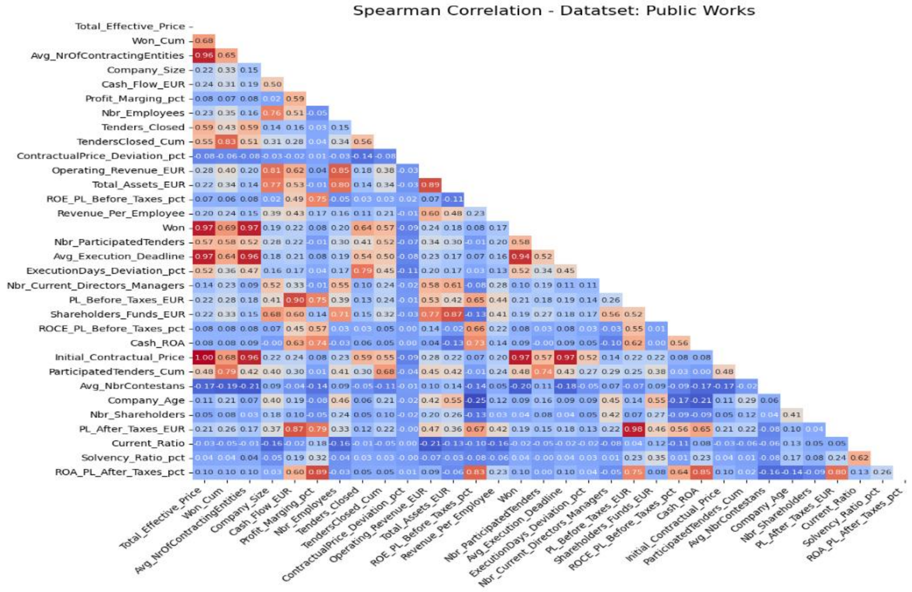
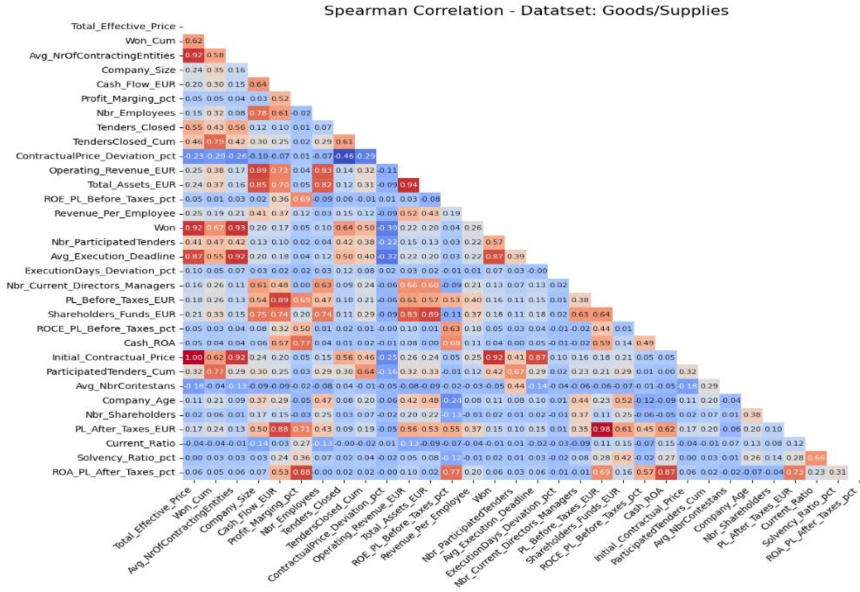


Figure 9 - Histogram with Distribution of Dependent Variable 'Won' of Public Works, Goods and Supplies and Services Datasets

Appendix L





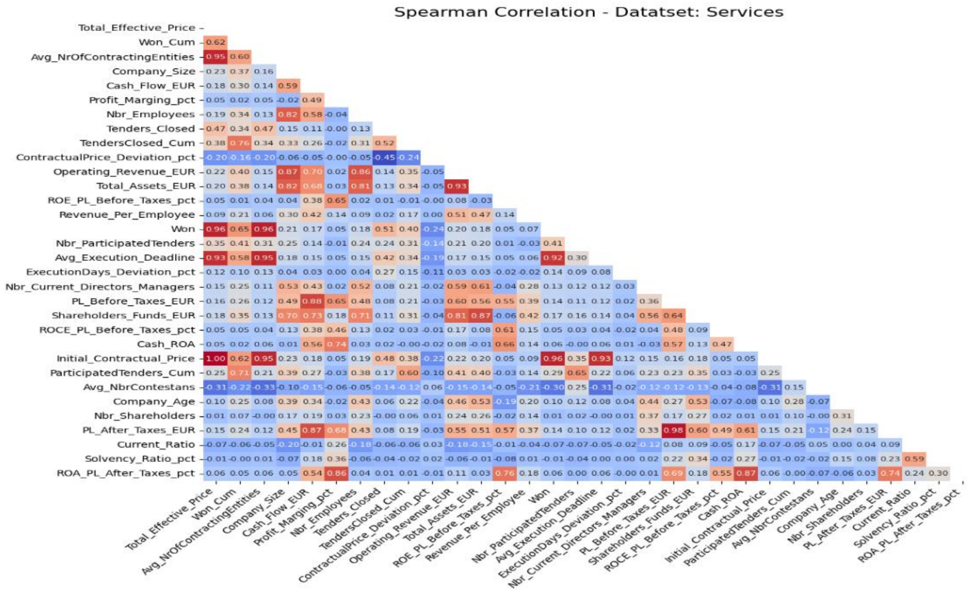


Figure 10 - Spearman Correlation of Public Works, Goods and Supplies and Services Datasets

Appendix M

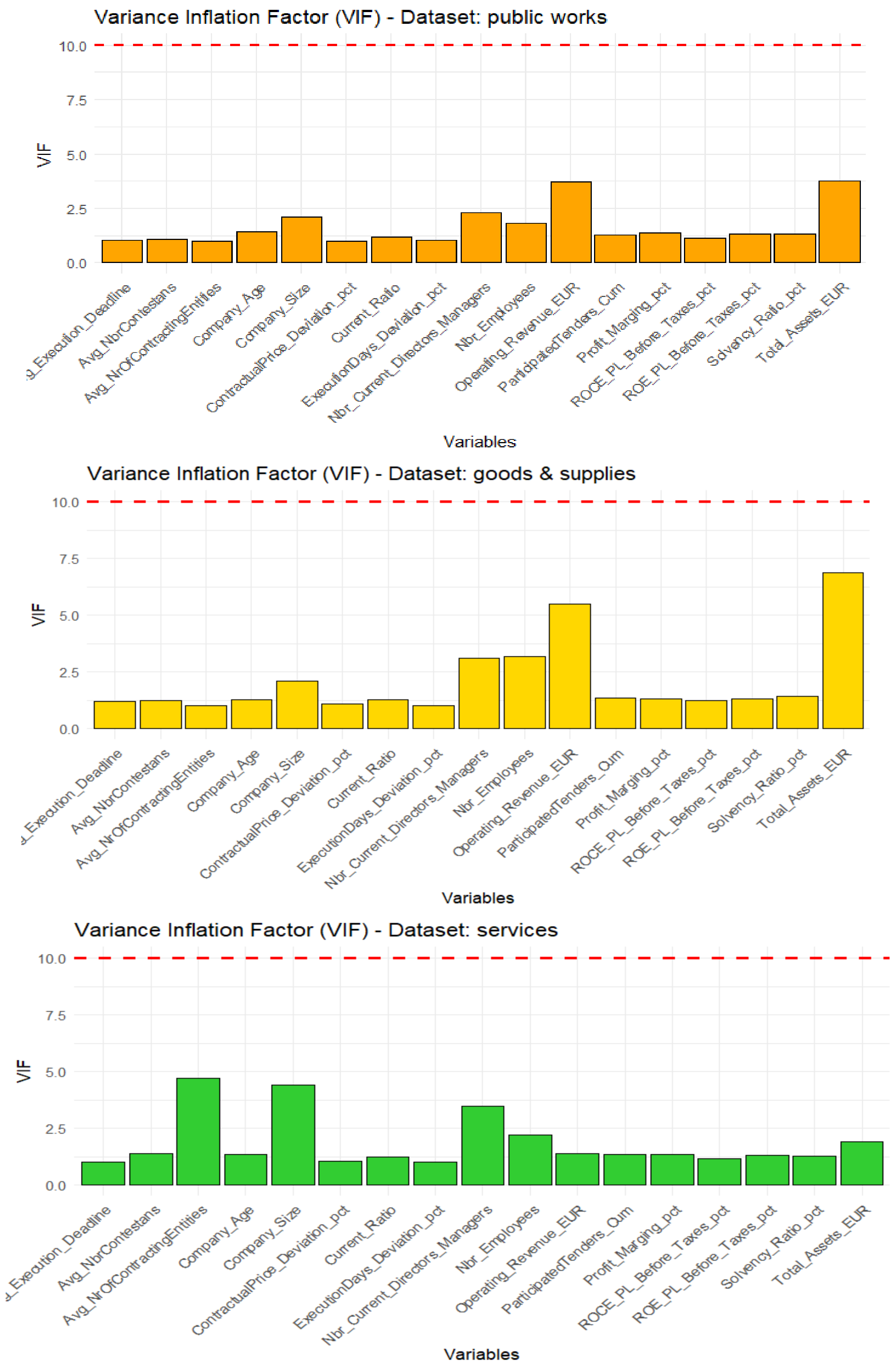


Figure 11 - Variance Inflation Factor of Dependent Variables in Public Works, Goods and Supplies and Services Datasets

Appendix N

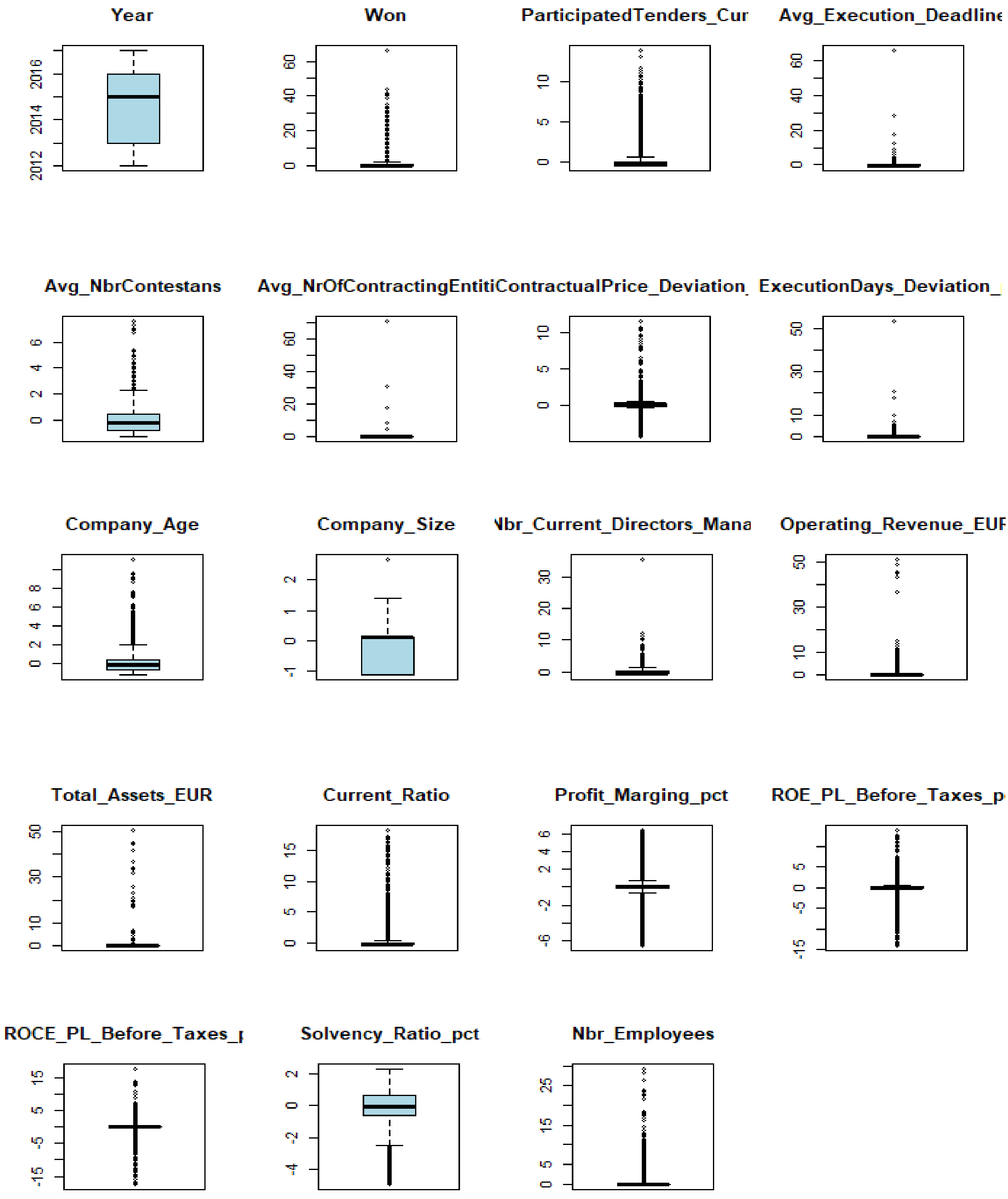


Figure 12 - Outliers Analysis for Variables in Consolidated Dataset

Appendix O

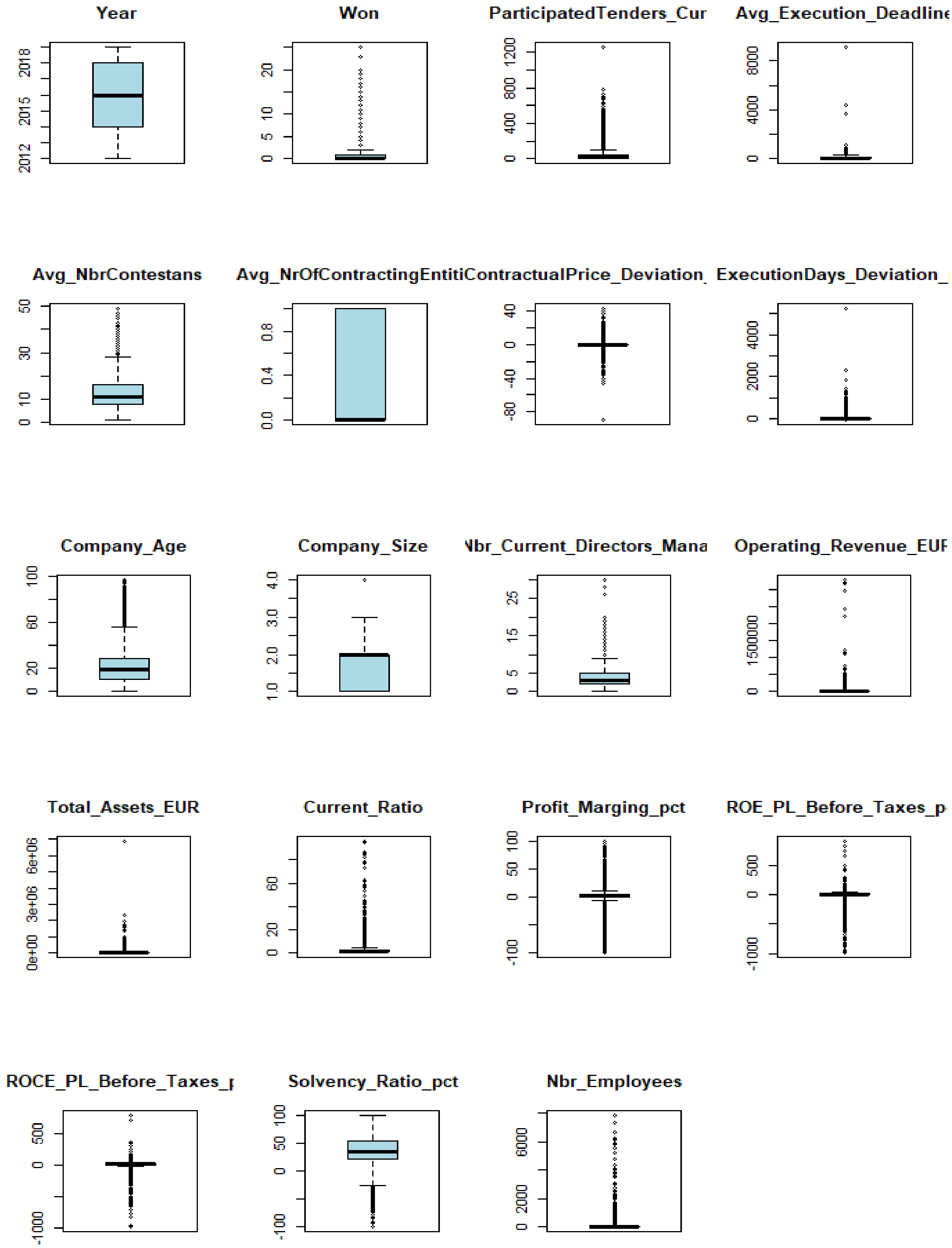


Figure 13 - Outliers Analysis for Variables in Public Works Dataset

Appendix P

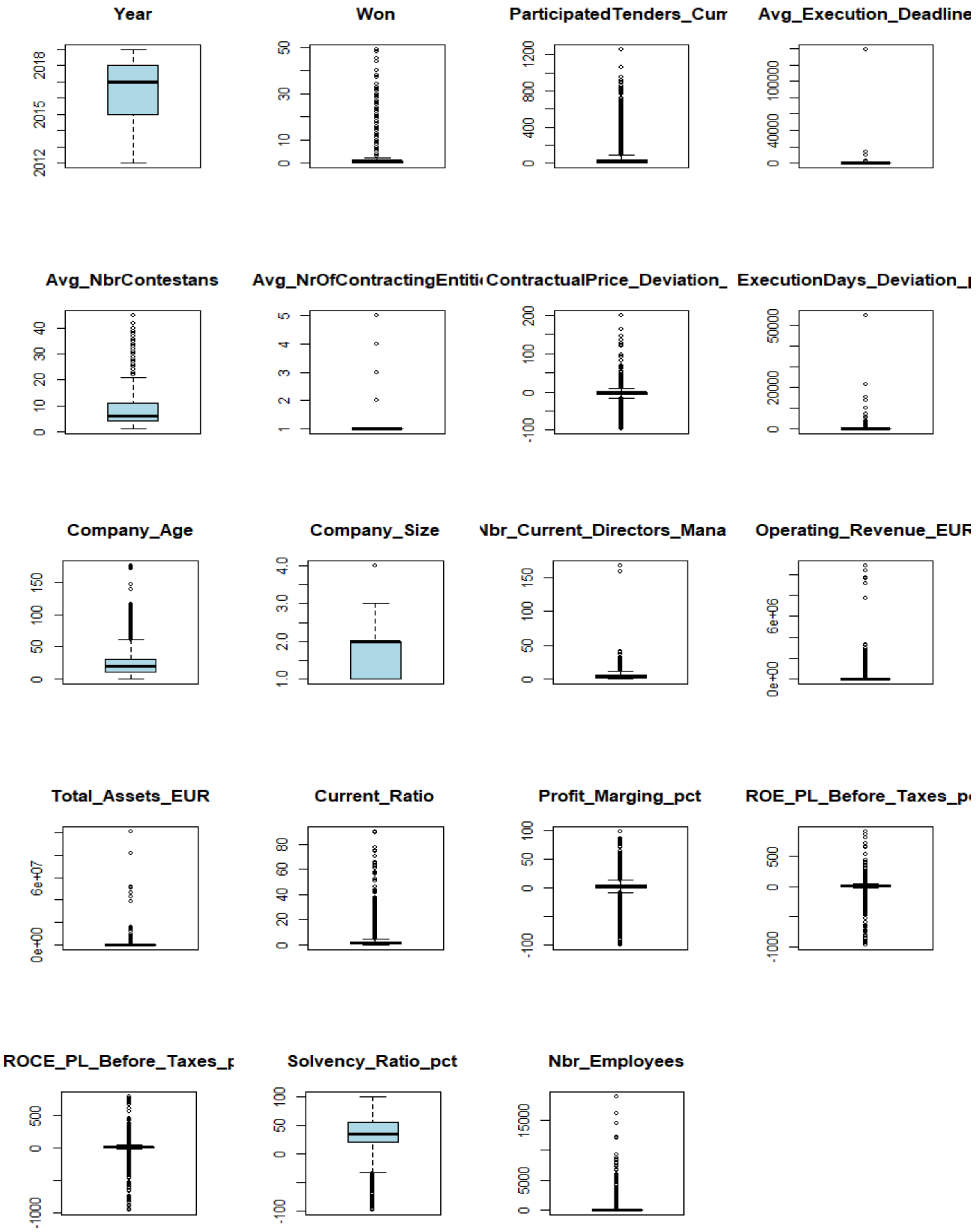


Figure 14 - Outliers Analysis for Variables in Goods and Supplies Dataset

Appendix Q

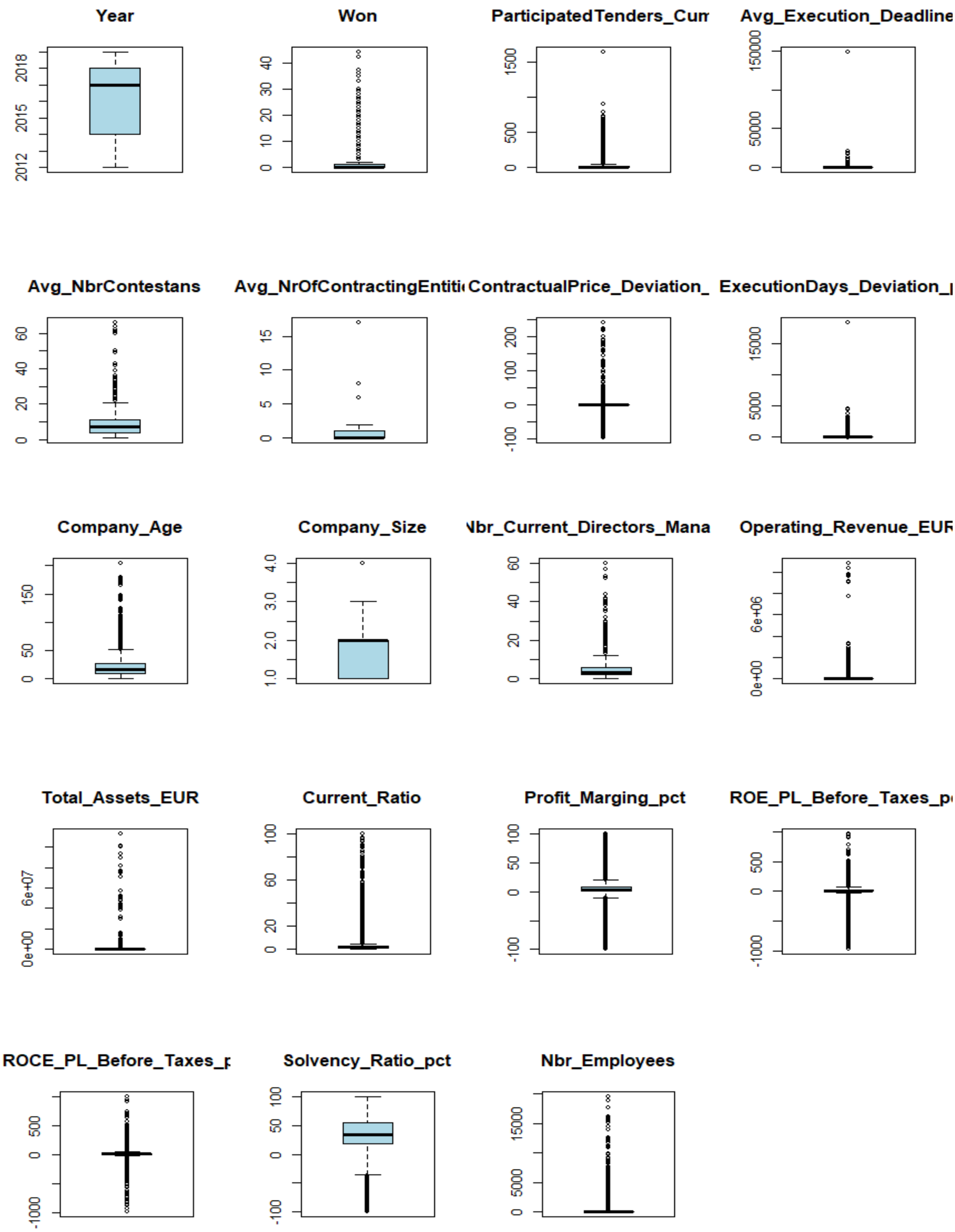


Figure 15 - Outliers Analysis for Variables in Services Dataset