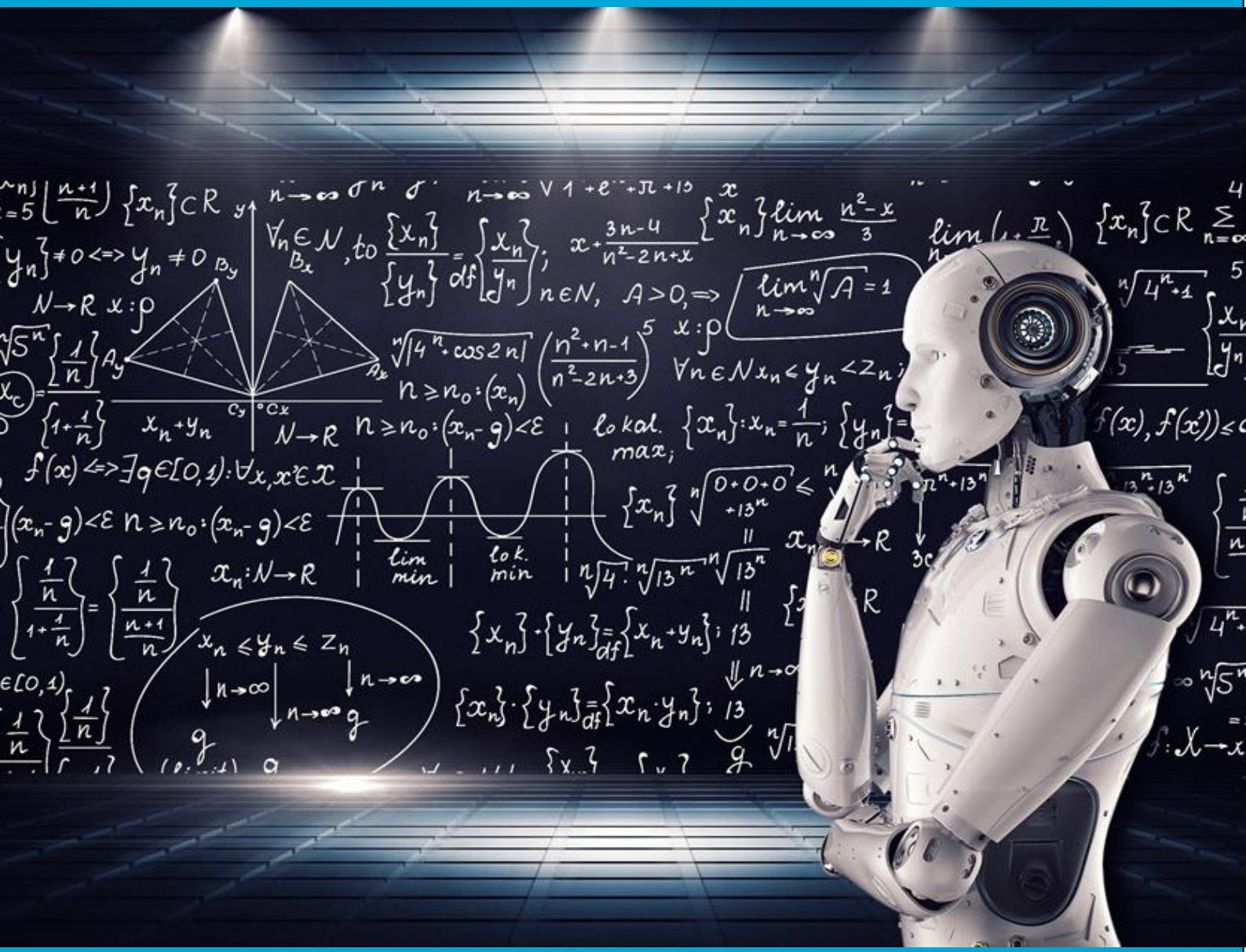
Tender Price Predictor

Predicting Tender Prices of Dutch Infrastructure Projects with Machine Learning

Master Thesis by B.C.J. Schleipfenbauer



TU Delft

bam

Tender Price Predictor

Master Thesis

By

B.C.J. Schleipfenbauer

Graduation Thesis:

University	Delft University of Technology
Faculty	Civil Engineering and Geosciences (CiTG)
Master program	MSc. Construction Management and Engineering

Graduation Committee:

Chair:	Prof. dr. Hans L.M. Bakker,	| TU Delft (CEG)
First supervisor:	Dr. ir. Marian G.C. Bosch-Rekveldt,	| TU Delft (CEG)
Second supervisor:	Ir. Jeroen Delfos,	| TU Delft (TPM)
Company supervisor:	Ir. Bas van de Weijer,	| BAM Infra

Preface

This thesis report 'Tender Price Predictor: Predicting Tender Prices of Dutch Infrastructure Projects with Machine Learning' has been written as the final component of the MSc. 'Construction Management and Engineering' programme of the Delft University of Technology.

Investigating how Machine Learning algorithms could be used to aid the construction industry was a natural choice for me, having obtained a bachelor degree in civil engineering at the University of Twente combined with a fascination for data. It was not possible to complete this project solely with my fascination for data and the academic and problem-solving skills developed at university. Therefore, I would like to thank everyone that made it possible to complete this research.

First, I would like to thank Bas van de Weijer and his colleagues of BAM Infra's risk management department for making it possible to graduate on this topic at BAM Infra. Bas van de Weijer made sure that I and my fellow graduate interns felt at ease in the organization, and helped me to get in touch with the right people for my research. Also, I would like to thank everyone that contributed to the research, both interviewees and colleagues from the risk management department.

Secondly, I would like to thank my supervisors of the Delft University of Technology for their input and the monitoring of my research process. The support and feedback from Marian Bosch-Rekveldt during the bi-monthly meetings have been valuable and highly appreciated. Furthermore, I would like to thank Jeroen Delfos for sharing his knowledge on the topic of Machine Learning and contributing to the development of my model. His help on my graduation research went beyond the expected input of a second supervisor. Also, I would like to thank Hans Bakker for his insights, feedback and for the supervision of my progress meetings.

Finally, I would like to thank my dear friends, family and roommates for the necessary distractions and support. Thank you for a listening ear when I had to complain when things were not going as planned, and for joining my celebrations when the model started working.

Bent Schleipfenbauer

Delft, January 2022

Executive Summary

The Dutch public procurement market is a multi-billion industry, with a total value of 73 billion euros per year. Competitive tendering is the most popular method of selecting a supplier for the required construction services. The realisation of a tender bid is an expensive and complex process, established on the intersection of various disciplines e.g. safety, constructability, finance, cost estimation and risk management.

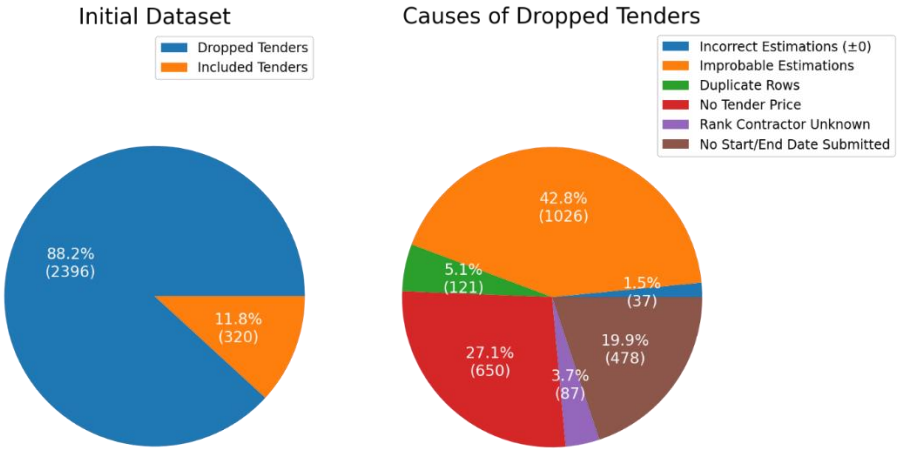
Machine Learning has been a popular method in various industries to predict future outcomes and uncover patterns in historical data but remains a rather novel phenomenon in the construction industry. Machine Learning models have been developed in the past to aid tender management, but not with a focus on predicting the contractor's tender price.

The objective of this research is to develop a Machine Learning tool that is able to predict the tender price of infrastructure projects accurately and is able to assist the contractor's tender professionals in their decision to tender. In order to achieve this objective, the following research question was formulated:

How can a Machine Learning algorithm, predicting the tender's price using tender project data, be developed to support the contractor's decision to tender?

The research framework, developed to answer the main research question, is based on the design methodology for new engineering systems (Roozenburg and Eekels 1995). First, a literature study is conducted to explore the state-of-the-art developments of Machine Learning within construction tender management, discover what the most popular regression algorithms are and what the most important tender features are that influence the tender price.

Based on the most important tender features and interviews with tender professionals of a Dutch contractor, an extensive list of 93 tender features is filtered until a final set of tender features remains. Data on these tender features are collected in order to be used as input for the Machine Learning model. After cleaning the raw dataset and applying outlier detection, only 222 tenders could be used as input for the ML model. This is less than 10% of the initial raw dataset, mainly caused by improbable estimations, no documented tender prices in the database or no initial start/end dates provided.



Causes Dropped Tenders, Source: Own Image

After obtaining and cleaning the tender dataset, three regression algorithms are developed with the purpose to predict the tender's price using the selected features. The main criteria used to select the algorithms are 'interpretability' and 'accuracy'. The models scoring best on these criteria are Linear Regression, Decision Tree Regression and Support Vector Regression. Linear Regression and DTR are

the easiest to interpret by deriving the feature coefficients, while SVRs tend to provide the highest accuracy while retaining the interpretability according to literature.

The SVR model performed the best with an R-Squared of 0.846, implying that 84.6% of the variance of the tender's price could be explained by the model. The SVR model includes an optimised set of features, which is a subset of the initial dataset. The initial estimate is considered to be significantly more important than the other features, as illustrated in the table below:

Optimised Subset of Features

Feature Name	Feature Importance
Estimate	0.866
Sqrt_Duration	0.051
Contract_RAW	0.053
Procurement_Price-Only	-0.088

Comparing the model's predictions to the actual tender price, a mean absolute percentage error (MAPE) of 23.5% is obtained which is equivalent to an accuracy of 76.5%. This value is marginally lower than the MAPE of the experts' estimations, which obtain a MAPE of 23.3 %. This appears to be caused by an incorrectly categorised datapoint. Modifying this datapoint improves the model's MAPE from 23.5 % to 22.0 %. Although the absolute mean deviation of the model's predictions is larger than the experts' estimations, the majority of the model's predictions are more accurate than the expert's estimations (53,6% vs 46,4%). Also, the average absolute price deviation in € of the model is lower than the experts' estimations (125.581 € vs 141.964 €).

During the validation interview, interviewees agreed that contractors may benefit from the Tender Price Predictor when it is possible to achieve predictions with a maximum error of 10% - 30 % of the winning tender price. The use-case of the Tender Price Predictor may be improved when project-specific or industry-specific characteristics would be used in order to meet the demands of more specialised industries within BAM. Also, the EMVI-component and complexity feature could be improved upon by including more quality-plan components and replacing the tender category feature with more specific complexity variables.

In order to implement the Tender Price Predictor in the organization of a Dutch contractor, attention should be paid to the required effort of the users and to ensuring a high quality of tender data. Requiring too much effort from tender managers entering the input data into the database may result in worse quality of data. Both the users of the model and the managers submitting tender data should be trained accordingly in order to obtain maximum effectiveness.

It should be noted that some important features, according to practice and literature, are omitted from the final dataset. Although the final set of features complies with the requirements of generic features and sufficient occurrences in either literature or the interviews, they were not present in the database of the contractor. These omitted features are: 'Project team experience' and 'Location'.

Invalid data and outlier detection resulted in a significant decrease (± -90%) of the useable dataset size. As a result, the ML models have used data on relative small-sized tenders due to large tenders not being represented in the dataset. In order to incorporate large tenders in the model's predictions, either missing data should be restored accordingly or larger tenders should be divided into smaller modules. Invalid data could have been checked manually by analysing project documents or by interviewing the responsible tender managers, but this has not been attempted due to time constraints.

Table of Content

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Meaning	Page
MSc	Master of Science	1
EU	European Union	1
PMBoK	Project Management Body of Knowledge'	2
AI	Artificial Intelligence	2
ML	Machine Learning	2
ANN	Artificial Neural Networks	4
EMVI	Economisch Meest Voordelige Inschrijving	5
SVM	Support Vector Machine	12
LASSO	Least Absolute Shrinkage and Selection Operator	14
SVR	Support Vector Regression	16
RBF	Radial Basis Function	16
DTR	Decision Tree Regression	17
OHE	One-Hot-Encoding	25
PCA	Principle Component Analysis	26
RFE	Recursive Feature Elimination	26
PC	PC	27
MAE	Mean Absolute Error	35
MSE	Mean Squared Error	36
RMSE	Root Mean Squared Error	36
MAPE	Mean Absolute Percentage Error	43
IV	Interviewee Validation	45

1. Introduction

This chapter provides an introduction to the graduation research's topic. This graduation research is conducted in order to complete the MSc 'Construction Management & Engineering' programme at the Delft University of Engineering. The function of the introduction is to familiarise the reader with the topic of the graduate's research, its objective and the relevance of the thesis.

1.1. Background Information

This section provides background information regarding tenders within the Dutch construction industry and the potential role of AI within tender management.

1.1.1 Forms of Procurement

Directive 2014/24/EU of the European Parliament on public procurement defines 'Procurement' as follows: "(Procurement) is the acquisition by means of a public contract of works, supplies or services by one or more contracting authorities from economic operators chosen by those contracting authorities, whether or not the works, supplies or services intended for a purpose." Procurement involves the organizational process of acquiring works, supplies or services between the contracting or client party and the supplier or contractor (Kerzner, 2003). The shape of the procurement process itself, the decision of executing project works internally or externally, depends on the make-or-buy analysis of the works, supplies, services or deliverables (Nicholas & Steyn, 2017). More detailed information on 'European Procurement Directives' and Dutch procurement law can be found in Appendices A&B.

The Dutch public procurement market is a multi-billion industry, with a total value of 73 billion € per year (Ministry of Economic Affairs and Climate Policy 2018). Clients, such as the Dutch government, have four ways to procure construction services (Winch 2020a): procure in-house, appoint a service supplier, launch a concours or issue an invitation for a competitive tender. Competitive tendering is the most popular method of selecting a supplier for the required construction services (Winch 2020a). In total, more than 80% of construction projects in the Netherlands are tendered competitively (Van de Rijt, Hompes, and Santema 2010).

The exact contents of competitive tendering activities depend on the project's aspects and the circumstances of the project (Agerberg 2012). One of the disadvantages of competitive tendering in the construction industry is the high tendering costs, estimated at 10% of the turnover (Winch 2020a). Insufficient time for the cost estimation process, poor analysis of cost data and lack of data processing techniques are among the most common causes of inaccurate tender cost estimating (Akintoye and Fitzgerald 2000). Depending on the contract's agreements between client and contractor and the nature of the incremental costs, either party may be accounted for the overrun resulting in smaller profit margins.

1.1.2 Cost Estimating within the Construction Industry

Project cost is one of the three important drivers of project success and together with time and quality forms the so-called 'Iron Triangle' of Project Management (Atkinson 1999; Pollack, Helm, and Adler 2018). The performance of a project on these success criteria denotes the degree of success or failure of a construction project. According to Flyvbjerg et al., 90% of all infrastructure projects exceed their budgets with an average overrun between 20-45%, depending on the purpose of the infrastructure (Flyvbjerg, Skamris, and Buhl 2004). Cost overruns may be caused by poor project descriptions or lack of scope, resulting in increasing costs eventually exceeding the initial budget (Nicholas & Steyn, 2017, p.282).

A contractor benefits from accurate tender estimating. If the contractor's tender team underestimates the tender's costs, the organization wins the tender but will end up with red figures due to higher actual costs than initially estimated. If the contractor's tender team would overestimate the tender's price, the organization would end up losing the tender indefinitely. An overview of the composition of a tender's estimated costs is given in figure 1.

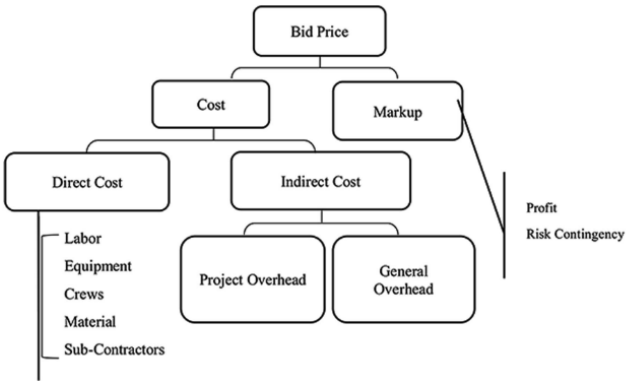


Figure 1 Tender Price Composition, Source: (Hashemi, Ebadati, and Kaur 2020)

Tender price estimating is an extensive process that could take up various months, with entire tender teams devoted to the tender process in order to acquire the project. Not winning the tender would result in losing potential turnover in the form of the contract's value and losing the investment in the tender team. After the contractor has considered aspects like project risk, tender costs, potential competitors, experience with both winning the tender as the construction of the project, a decision-to-bid is made (Brook 2008).

Cost estimating is defined by the PMBoK as the "process of developing an approximation of the cost of resources needed to complete project work" ("PMBOK 6th Edition," 2017, p.206). These estimations are made to determine what the budget is for a project and can be used to measure the actual performance of a project by comparing the actual costs to the estimated costs (Nicholas & Steyn, 2017, p.282). Traditional cost estimations are made by calculations of cost engineers, either by analogous estimating or parametric estimating. Analogous estimating is the process of projecting a future project's costs by direct comparison to similar projects ("PMBOK 6th Edition," 2017, Ch. 7.2.2.2). Parametric estimates are based on mathematical relationships between projects' parameters (Nicholas & Steyn, 2017). Mathematical relationships are derived from multivariate regression methods or extensive data analyses.

1.1.3 Machine Learning and Opportunities for Tender Management

With the large technological advances of the past decades, forms of AI are blossoming in countless disciplines within the scientific field and in business practices. For example, applications of Machine Learning can be found in many different scientific domains including bioinformatics, computer science, statistics, surveillance, speech recognition and as tool in many data-intensive disciplines like finance and astronomy (Baştanlar and Özuysal 2014; Emerson et al. 2019; Mitchell 1997; 2006). Seemingly, the construction industry lags behind these industries with a lack of data-driven Machine Learning developments (Gondia et al. 2020; Hussain et al. 2018).

The foundation of ML and AI in general originates from the late fifties and early sixties of the previous century. Frank Rosenblatt programmed a machine that recognised the different letters of the alphabet (Fradkov 2020). Rosenblatt's aim was to mimic the human brain or human intelligence perceptron. The work of Rosenblatt initiated the global boom of AI development by inspiring scientists, professionals and computer enthusiasts to develop their own imitations of human intelligence (Haykin 2009).

ML algorithms are a form of AI which uses 'simulated experiences' to make predictions based on historical data. The objective of ML is defined as the following:

"A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P,* if its performance at tasks in *T*, as measured by *P*, improves with experience *E*."(Mitchell 1997).

Given this underlying relationship, ML is defined as a subset of AI (figure 2).

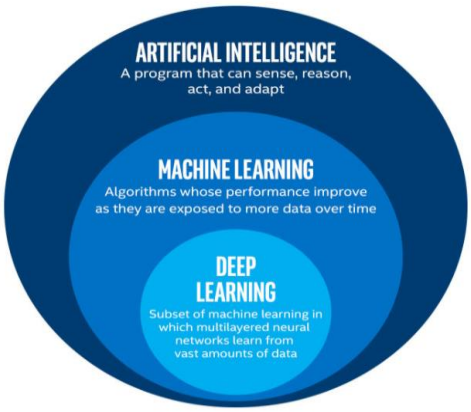


Figure 2 Relationship AI and Machine Learning, Source: (Singh, 2018)

Within ML, three main types of learning can be distinguished: 'Unsupervised Learning', 'Reinforcement Learning' and 'Supervised Learning'. The purpose of unsupervised learning is to find structure and relationships behind data classes without knowing the context of the data entries (Jain, Murty, and Flynn 1999), while reinforcement Learning does not concern the unravelling of structures within data but concerns maximisation of an agent's rewards by means of trial and error (Sutton and Barto 2014). These types of ML do not fit the problem of tender price estimating as well as supervised learning does. This research focuses on the application of supervised learning, namely regression analysis. Regression analyses are used for numerical, labelled outputs (Metwalli 2020) such as the tender's price.

The objective of supervised learning is to analyse data in order to simulate relationships and dependencies between input features and output variables (Fumo 2017). Supervised learning is characterised by the labelling of data entries which are categorized in advance by knowledgeable human experts, therefore the adjective 'Supervised'.

Supervised Learning algorithms can be used to solve two types of problems: regression problems or classification problems. Regression problems may be solved to predict real-valued outcomes based on underlying interrelationships. Classification problems relate to the prediction of a discrete output based on a collection of input data.

Historical data on previous construction projects may be of use to predict the tender's price to gain an edge in the market. Traditionally, data is used implicitly in the form of an expert's experience, but this may also be employed by ML algorithms to support the contractor in their decision-to-bid based on previous performances on tenders.

1.2. Research Gap

Several studies noticed that the construction industry lags behind, compared to other scientific and business disciplines, with respect to the development of AI or ML (Seidu et al. 2020; Gondia et al. 2020; Hussain et al. 2018). However, some models have been developed to aid cost estimating within the construction industry.

ML models have been developed in the past to predict prices within the construction industry but with a different focus. For example, ML models have been designed with the objective to aid either the client by evaluating submitted bids (Zhang, Luo, and He 2015) or to support subcontracted engineering consultants to provide the client with a quotation (Matel et al. 2019). Using respectively regression analysis and ANNs,

bids were predicted for the corresponding parties. The contractor's perspective has not been represented in these studies.

Some predictive models have been designed for contractors in the infrastructure construction industry, but these do not focus on the tender's price. The model designed by Kultin and his co-authors attempted to predict the winning probability of the contractor (Kultin, Kultin, and Bauer 2021). The ANN model designed by Elhag and Boussabaine aim to predict the lowest price i.e. winning tender price (Elhag and Boussabaine 1998). The model, however, does not focus on the infrastructure industry but focuses on the building of schools.

It can be concluded from state-of-the-art developments that ML models have been applied in case studies with success. It remains unknown how effective similar models are in predicting tender prices of infrastructure projects and how these could be implemented in the contractor's tender business.

1.3. Problem Statement

The realisation of a tender bid is an expensive and complex process, established on the intersection of various disciplines e.g. safety, constructability, finance, cost estimation and risk management. While ML has been a hot topic within science and engineering during the 21st century, it remains a rather novel phenomenon in the construction- and quantity surveying industries (Seidu et al. 2020; Gondia et al. 2020; Hussain et al. 2018).

The tender characteristics that can be collected are endless. This is a result of the multidisciplinary character, size and complexity of construction projects. What the most influential variables are that determine the tender's price is unclear.

Machine Learning remains a relatively unknown topic within the contractor's tender industry while its application may potentially improve the contractor's decision-making regarding tender bids.

1.4. Research Objective

To develop a Machine Learning tool that is able to predict the tender price of infrastructure projects accurately and is able to assist the contractor's tender professionals in their decision to tender.

1.5. Research Question

How can a Machine Learning algorithm, predicting the tender's price using tender project data, be developed to support the contractor's decision to tender?

Subquestions (SQ)

1) What tender price features influence the tender's price?
2) What Machine Learning algorithms are most suitable, taking into account the available data of the contractor?
3) How accurate are tender price predictions by applying Machine Learning algorithms using historical project data?
4) How can the Tender Price Predictor effectively be used within Dutch tender practices?

1.6. Research Scope

The following sections will explain the scope of this research by introducing the graduation company and its previous attempt of applying ML within tender management.

1.6.1. Graduation Company: Dutch Contractor

The graduation thesis is written in collaboration with the 'risk management'-department of a Dutch contractor. The contractor, one of the largest European contracting firms active in construction and civil engineering with operations across the world, facilitates knowledge and data in order to design and train the Tender Price Predictor.

1.6.2. Implementation Phase ML Model

Three main steps exist within competitive tendering: preparation of documents, the start of the tender phase and the execution of works or services. The tender phase itself can be divided into parts as well, and differ from contractor to contractor. The contractor has divided its tender procedure into the following so-called 'Stage Gates', illustrated in figure 3. Each stage-gate has its own specific activities and documents which are required to submit according to internal policies.

Figure 3 Tender Phasing Contractor, Source: BAM Infra

The ML model is designed to support the contractor in its decision to tender. Early implementation can prevent the waste of money as a result of bid preparation on tenders that the contractor has not performed well on historically. If the contractor does not decide to make a bid, corresponding tender costs and design costs are saved. The view on this choice of this implementation phase is shared by employees of the Dutch contractor during informal meetings.

After deciding whether to tender or not, much more time and costs are invested by the company and its employees to win the tender. Depending on the size and complexity of the tender this may take multiple months of exploring the risks and opportunities, costs, bills of quantities and planning.

1.6.3. Initial Attempt Tender Price Predictor Contractor

The research contributes to contractor's practices by the development of the Tender Price Predictor. Tender data is used in order to train the algorithm, with the aim to assist the contractor's tender professionals in their decision-making processes.

In 2019, an ML model was designed by the contractor to make predictions of the tender's price, discounted tender prices and EMVI-scores for tenders larger than 30.000.000 €. The feature used to make these predictions were solely based on the ceiling price as set by the client. Data on 27 tenders were used as input for the model.

The initial attempt by the contractor is used as inspiration for this research topic. The focus of this study is to expand the number of features used, beyond the ceiling price, increase the dataset and create an ML model that can be used as a tool by tender professionals themselves to make predictions of the tender's price.

1.7. Research Relevance

ML algorithm-based tender price predictors by means of multivariate regression or classification are not new. Various studies have attempted to predict prices within (construction) procurement by means of ML (Zhang, Luo, and He 2015; Kultin, Kultin, and Bauer 2021; Matel et al. 2019; Wang, Yu, and Chan 2012; Stiti and Yape 2019).

Where some authors focus on the probability of winning the tender (Wang, Yu, and Chan 2012; Kultin, Kultin, and Bauer 2021) or the forecasting of tender bids or construction prices (Stiti and Yape 2019; Matel et al. 2019; Zhang, Luo, and He 2015), this study differentiates itself from the latter by focusing on infrastructure contractor's perspective of predicting tender prices.

1.8. Reading Guide

Chapter 2 presents the methodology of the research study.

Chapter 3 provides the literature study of this research. The purpose of the literature study is to become more familiar with relevant topics.

Chapter 4 contains the selection process of tender features for the ML models' input. The most important tender features are identified through literature studies and interviews with tender professionals. Subquestions 1 is answered in this chapter.

Chapter 5 explores the preprocessing phase of the data and the steps that are undertaken in order to clean the raw data. Based on the literature study of Chapter 3 and the available tender data, three algorithms are selected as potential Tender Price Predictor candidates. Subquestion 2 is answered in this chapter.

Chapter 6 revolves around the development of the Tender Price Predictor. An overview of the separate development steps is provided. Furthermore, the performance of the ML algorithms is compared to select the optimal Tender Price Predictor. Subquestion 3 is answered in this chapter.

Chapter 7 includes the evaluation of the designed Tender Price Predictor by means of interviews with tender professionals. The purpose of this chapter is to evaluate the potential impact of the Tender Price Predictor, and investigate how useful such an ML model may be in Dutch Tender Practice. Subquestion 4 is answered in this chapter

Chapter 8 consists of the discussion and limitations.

Chapter 9 contains the conclusion and recommendations of the research. The research question is answered in this chapter.

2. Research Methodology

The research design aims to answer the main research question by modularly breaking it down into smaller parts. The research methodology is based on the development method for new engineering systems of Roozenburg and Eekels (Roozenburg and Eekels 1995). The methodology is presented in figure 4 and elaborated below.

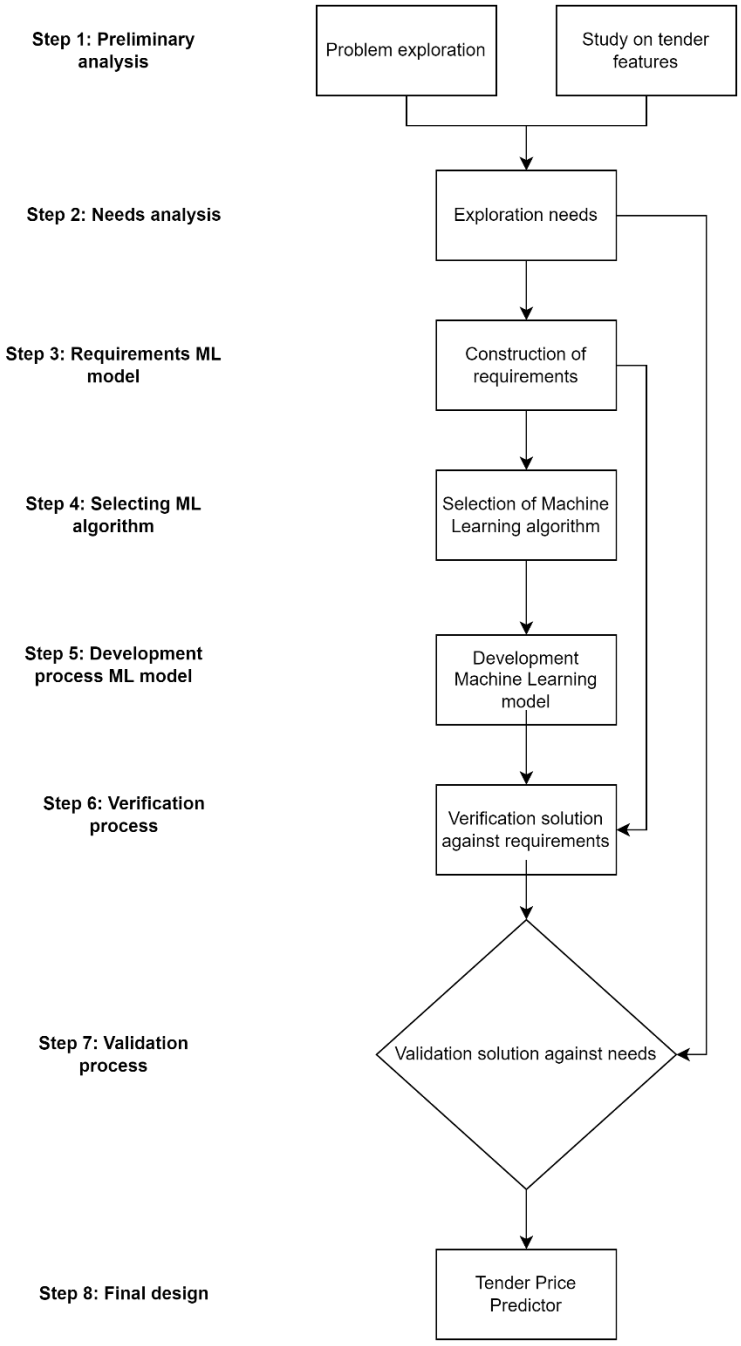


Figure 4 Research Methodology 'Tender Price Predictor ', Source: Own Image

2.1. Step 1: Preliminary Analysis

The preliminary analysis consists of a background study on both the problem and on tender price influencing features. The objective of this analysis is to improve the understanding of what the engineering system, the Tender Price Predictor, is designed for. A combination of literature study and expert interviews is conducted to collect data on the most important tender features and to better understand their influences on the tender price.

Semi-structured interviews are conducted to benefit from the advantages of both unstructured as structured interviews. According to the study of Alsaawi (2016), semi-structured interviews are most suitable for interviews where the researcher already has an overview of the to-be-discussed subjects but where the depth of the interview is not restricted by the format as the interviewees can extend the scope of the topic (Alsaawi, 2016, p.151). This approach is fitting, given that the interviews are a follow-up of the literature study on the tender features.

The intended result of the preliminary analysis is an ordered overview of the most important tender features according to both literature and practice. The purpose of the collected tender price features is a proposal of the ideal variables to use as input for the ML model.

2.2. Step 2: Needs Analysis

An exploration of needs is conducted to explore the exact needs of the potential users, the tender professionals. Needs analyses are used to investigate if an operational need exists, and whether the conceptualised engineering system fulfils this gap (Kossiakoff et al., 2011, p.139).

The needs are based on the findings in the first round of semi-structured interviews. The intended result of the needs analysis is an overview of the needs of the user. With the needs of the user clear, it is possible to construct the requirements of the ML model.

2.3. Step 3: Definition of Requirements of the ML Model

The needs analysis makes it possible to define the requirements of the ML model. The requirements of the ML model have the purpose transform the needs into tangible criteria of the model's performance.

2.4. Step 4: Choice of ML Model

In order to select the most appropriate ML model, an analysis of the strengths and weaknesses of popular regression techniques is conducted. The most suitable ML algorithms, based on the strengths and weaknesses, requirements and the type of data available are developed into Tender Price Predictors.

2.5. Step 5: Development Process of the ML Model

It is possible to start the development process of the ML model with the set of most suitable algorithms determined and the tender data collected. The development of an ML model is an extensive, iterative process with various loops between the development of the model and the dataset. The performances of the ML models are evaluated to determine which model is most suitable for the prediction of tender prices.

The model will be programmed using Python, based on prior knowledge and coding documentation found online. Within the SciKit-Learn, a popular library of Python containing various state-of-the-art learning models, it is possible to compare the performance of separate ML algorithms rather quick and easy with much less code required compared to non-dedicated libraries (Brownlee 2020a) (Pedregosa et al. 2011).

2.6. Step 6: Verification Process

With the ML model designed, it is checked whether the previously set requirements of the model have been met. The model's design should be modified accordingly if the output of the model does not meet

the requirements which initiate a new loop until all requirements are satisfied, or this cannot be accomplished at all.

2.7. Step 7: Validation Process

The validation process is the final component of the Tender Price Predictor design. In the validation process, the model's performance and functioning are validated by means of a set of interviews with Dutch tender professionals. The interviews are conducted to check if the Tender Price Predictor model could fit in construction tender management practices and whether tender managers would use such algorithmic tools. The purpose of the second round of interviews is to investigate how usable the Tender Price Predictor is in practice.

3. Literature Review

The purpose of this chapter is to improve the understanding of subjects relevant to the development of the Tender Price Predictor. The following subjects are studied in the literature review:

- Section 3.1 explains the competitive tendering procedures in the Netherlands.
- Section 3.2 introduces the most tender price influencing features, according to literature.
- Section 3.3 explores the state-of-the-art ML applications within Tender Management.
- Section 3.4 introduces the most popular regression techniques, the math behind the algorithms and how these interact with the data.
- Section 3.5 provides the conclusion.

Section 3.2 and 3.3 have another purpose besides familiarizing with tender price features. The literature study on tender price features is used together with the interview results of section 4.2 to identify which features to use in the ML model.

3.1. Competitive Tendering in the Netherlands

Within competitive tendering, the client party issues a description with selection criteria and requirements and the contracting party is selected on basis of their submitted description of the to-be-constructed works, supplies or goods (Winch 2020b). The tender proposal describes the project plan of what is to be done during the project's contract timespan e.g. what works and activities are executed, planning of the described activities, an estimation of the costs and a risk management plan (Nicholas & Steyn, 2017, p. 132). Competitive tendering is characterised by its transparent selection criteria and open market competition and therefore stimulating production efficiency (Winch 2020b). Different national and European procurement procedures are possible for the contracting authority to follow.

Contracting authorities should take into account the following aspects for every contract in order to decide on which procurement procedure to follow (Chao-Duivis 2019):

1. Size of contract
2. Transaction costs for contracting authority and tenderers
3. Number of potential tenderers
4. Desired result
5. The complexity of the contract
6. Type of contract

Public procurement by means of competitive tendering consists of three steps. The first step is the preparation of the assignment of works or services (PIANOO 2021a). During this part of the procurement process, the client determines what its desires are, what European and national rules apply and draws up the description of to-be-delivered works or services.

The second step entails the start of the tender phase. The tender phase commences when the contracting authority announces the tender to the market and potential tenderers sign-up for the tender (PIANOO 2021b). The credentials and references of the tenderers are checked to verify whether the candidates satisfy the suitability requirements.

The final step of the public procurement process is the execution of the works or the delivery of services as described in the contract (PIANOO 2021c). The contractor and contracting authority both comply with the agreements made in the contract. The contracting party or client provides details on the works and location, the form of contract and specific requirements on the works and services to be executed by the potential tenderers (Brook, 2008, p.51).

Specification of the award criteria should be provided to the potential tenderers together with the tender details. The winner of the tender is the bid that scores best according to the award criteria. The tender winner selection procedure is illustrated in figure 5 on the next page (Dreschler 2009).

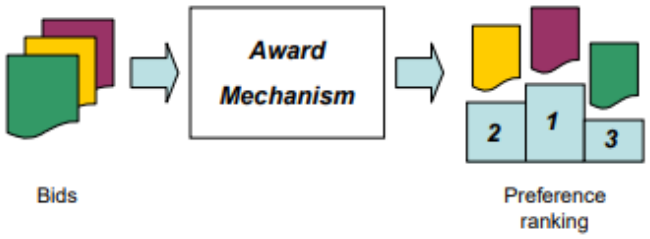


Figure 5 Tender Award Mechanism, Source: (Dreschler, 2009)

Three award mechanisms are commonly accepted within the Netherlands, all following the procurement principles (Overheid.nl, n.d., Article 2.114):

1. Best Price-Quality Ratio
2. Lowest Costs Based on Cost-Effectiveness
3. Lowest Price

Best Price-Quality Ratio is better known in the Netherlands as 'Economisch Meest Voordelige Inschrijving' or EMVI. The EMVI method is the most common procurement procedure for large tenders such as infrastructure projects (Overheid.nl 2012). Both the quality and the price of the tender play a role in EMVI in order to select the winning bid. Potential tenderers may achieve 'fictional discounts' when the description of works perform well on criteria like Sustainability, Nuisance and many more. The fictional discount for the specific tender component may be larger than the added costs of the component, resulting in a smaller 'final' tender price.

3.2. Tender Price Features in Construction Tender Estimating

This section is devoted to the literature study of exploring what tender price estimating features are the most influential in construction tender estimating according to the classical price estimating industry.

Odusami & Onukwubu devoted their study to the accuracy of cost estimates (compared to the lowest acceptable tender of the tender competition) in Nigeria, and what factors are most influential (Odusami and Onukwube 2008). The authors distinguish responses on the respective years of experience of the responders on their 6-page questionnaire (n=50), dividing the respondents into groups of 'less experienced', 'moderately experienced' and 'most experienced' respondents

Elhag et al. (2005) identified 67 features, grouped into six different categories, that play a role in building cost models of both quantity surveyors as contractors (Elhag, Boussabaine, and Ballal 2005). The features have been collected by means of a survey study across the quantity surveying discipline in the United Kingdom, with the surveyors ranking the separate features in order of importance. These features were divided into six different categories: Client characteristics, Consultant/design parameters, Contractor attributes, Project characteristics, Contract procedure/procurement method and Market conditions.

Akintoye (1999) conducted an empirical study on cost estimating influencing factors of construction contractors in the United Kingdom (A. S. Akintoye 1999). The author identified the 23 most influential factors among a large number of influencing factors by means of a survey sent to small, medium and large contractors based on annual turnover.

Elhag (2002) concluded that within its regression model a strong relationship exists between the tender's price and the gross area, no specific relationship between tender price and the total duration (Elhag, 2002, p.267). Interestingly, when comparing multiple estimation models, both regression analyses and ANNs,

Elhag discovered that models using two cost factors (duration and gross floor area) are more accurate than models utilising 13 cost factors (Elhag, 2002, p.269-270).

Combining the results of the tender price feature study results in an extensive list of 77 different features. An overview of these features can be found in Appendix E.1.

3.3. ML within Tender Management

This section is devoted to the literature study of exploring which ML algorithms have been applied in construction tender management in the past. The used input features by the authors are presented in Appendix E.2.

Zhang et al. (2015) used ML algorithms to evaluate construction element pricings during the bid submission (Zhang, Luo, and He 2015). The result is a pricing range for the construction items discussed in the conducted system analysis. The authors identified features like excavation depth, shape, hydrological conditions, soil conditions and surrounding environment as important project features. The reasoning for the particular set of project data is not provided.

Elhag and Boussabaine (1998) produced two ANN construction cost estimation models which are designed to predict the lowest tender price of schools (Elhag and Boussabaine 1998). Model II uses 13 selected cost determinant factors whereas model I only applies 4 input features: type of building, gross floor area, number of stories and project duration. The features used are numerical or categorized features. Model I and model II had corresponding accuracies of 79.3% and 82.2%. The authors concluded that "The more significant factors contributed in developing an ANN model, the better the outcomes" (Elhag and Boussabaine 1998, p.226) In total 30 projects were used in the dataset.

A recent study by Kultin et al. describes how a binary classification algorithm analyses the probability of successfully winning the tender for a construction project (Kultin, Kultin, and Bauer 2021). The authors compared the performance of two models, logistic regression and SVMs. SVMs are supervised learning algorithms that can be used to identify relations in large datasets and have a high generalization ability (Cortes and Vapnik 1995). SVM algorithms classify data by determining which data points are the supporting vectors within the training data set, and form the hyperplane or distinction within the dataset to classify the different data classes (Cheng and Wu 2009).

The algorithms designed by Kultin et al. assign a binary value to each project, either "prospective" (1) or "unpromising" (0), based on the project's performance on corresponding Tender features (Kultin, Kultin, and Bauer 2021). A set of 11 attributes is used to assess the probability of winning the tender. Examples are 'Type of Work', 'Preliminary budget' and 'Number of Contracts Accepted by the Manager'. This case shows that ML algorithms can be used in complex cases like the estimation of tender prices, but a substantiation for the Tender features selection is not provided by the authors.

An ML model was developed by Matel et al. to predict the total cost of engineering services provided by an engineering consultancy firm (Matel et al. 2019). Based on information collected during the tender phase, the authors discovered that the features that were thought to be most influential according to the expert panel differed from the outcome of the model. Although the provided rankings of the authors differ from the ranking of the model, the features collected through the interviews are not of lesser importance as a result. The most relevant tender features for cost estimating according to the model are: 1) Intensity (average hours spend per week per project team member), 2) Number of project team members, 3) Project duration, 4) Collaboration disciplines, 5) Contract type, 6) Project phases and 7) scale of work. An ANN model was used to achieve the engineering services' costs, with an average correlation coefficient of 0.99 between the inputs and the test set while achieving a MAPE of 13.65%.

A model that estimates both the construction costs and the scheduling success was designed by Wang et al. using neural networks and SVM to compare the corresponding efficiencies (Wang, Yu, and Chan

2012). The SVM model was proven to be more accurate (92%) than its neural network counterpart (80%). The author collected data by scoring 92 construction projects for 64 separate elements on a 0-5 scale.

3.4. Popular Regression Algorithms

This section is devoted to the introduction of popular regression techniques in ML. A short description of the algorithm, its inner workings and its ideal properties are provided below. The purpose of this section is to become more familiar with popular regression algorithms used in ML.

3.4.1. Linear Regression

Linear regression is the most simple form of parametric regression and is widely used to predict continuous outputs, even in cases when the problem is non-linear (Bonaccorso 2017). Linear regression attempts to fit independent input variables X by means of a linear model. Within linear regression, it is assumed that a linear correlation exists between the independent variables X and the dependent variable Y. In the equation below, the predicted dependent variable ŷ depends on the linear combination of set X and the coefficient factors α_i. The equation for a linear model is shown in equation 5.

$$\hat{y} = \alpha_0 + \sum_{i=1}^{m} \alpha_i x_i + \varepsilon_i, \{\alpha_0, \alpha_1, ..., \alpha_m\} \in R \quad (1)$$

By minimising the error of the equation, the ideal linear equation may be found which approximates the data the most (Pant 2019). It performs best when linear relationships exist between the various features (Elite Data Science 2019).

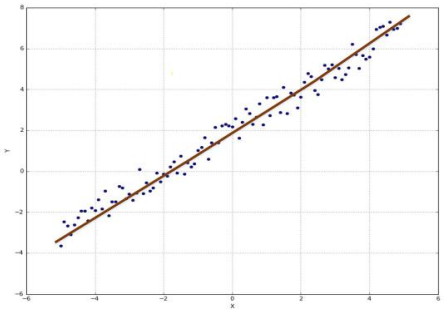


Figure 6 Linear Regression in 2 dimensions, Source: (Bonaccorso, 2017, p.129)

Advantages

- Simple and easy to interpret, especially when visualised in two dimensions like in figure 3.
- Coefficients of each feature may be acquired, providing insights into the sensitivities of the model.

Disadvantages

- Lacks the ability to make relationships of more complex problems or non-linear relationships (Elite Data Science 2021). Non-linear features should be re-engineered into linear features.
- Input data should be independent, and the algorithm's sensitivity to outliers (Flom 2018).

3.4.2. Ridge Regression

Ridge regression is a modified linear regression algorithm. The purpose of this algorithm is to normalize the weights of the features to prevent the overfitting of the model. As explained above, linear regression

aims to minimise the error or 'cost function' of the algorithm. The equation for the cost function can be found in equation 2 (Bhattacharyya 2018a).

$$\sum_{i=1}^{m}(y_i - \hat{y}_i) = \sum_{i=1}^{m}(y_i - (\sum_{j=0}^{p}\alpha_j * x_{ij}))^2 \quad (2)$$

Ridge Regression adds an extra term to the cost function, an extra penalty, which decreases the value of coefficients α_i or α_j. The modified 'ridge cost function', which is to be minimized in ridge regression, is illustrated in equation 7.

$$\sum_{i=1}^{m}(y_i - \hat{y}_i) = \sum_{i=1}^{m}(y_i - (\sum_{j=0}^{p}\alpha_j * x_{ij}))^2 + \lambda\sum_{j=0}^{p}\alpha_j{}^2 \quad (3)$$

Given that equation 3 is to be minimised and penalty term $\lambda > 0$, a linear equation with smaller coefficients is opted over equations with larger coefficients. This is called L1 regularization, the penalty function which characterizes Ridge Regression.

$$\text{For } \lambda \approx 0: \sum_{i=1}^{m}(y_i - \hat{y}_i) = \sum_{i=1}^{m}(y_i - (\sum_{j=0}^{p}\alpha_j * x_{ij}))^2 + 0\sum_{j=0}^{p}\alpha_j{}^2 = \sum_{i=1}^{m}(y_i - (\sum_{j=0}^{p}\alpha_j * x_{ij}))^2 \quad (4)$$

Advantages

- Ridge regression can be used when many input features are used when the features are strongly correlated while preventing overfitting at the same time (Sneiderman 2020).

Disadvantages

- Increased bias.
- Complex and harder to interpret than ordinary linear regression.
- While linear regression tends to underfit data, polynomial regression has the tendency to overfit the data when the dimension of the regression line increases. This means that the equation has low bias, fitting the problem's data well, but has high variance as the equation will only fit the current training set well. When data is implemented from outside the training set, for example, a new testing (data)set, it would most likely not lay on the overfitting equation.

3.4.3. Lasso Regression

The LASSO regression algorithm functions like ridge regression but uses a different penalty function. LASSO minimizes the absolute magnitude of coefficients instead of squaring them (Bhattacharyya 2018b). The cost function of LASSO regression can be found in equation 5.

$$\sum_{i=1}^{m}(y_i - \hat{y}_i) = \sum_{i=1}^{m}(y_i - (\sum_{j=0}^{p}\alpha_j * x_{ij}))^2 + \lambda\sum_{j=0}^{p}|\alpha_j| \quad (5)$$

The addition of this penalty statement is called L2 regularization. L2 differs from L1 in that some feature coefficients may converge to zero resulting in lesser important features being left out of the regression equation (Taunk 2020).

Advantages

- LASSO regression can be used when many input features are used when the features are strongly correlated while preventing overfitting at the same time (Sneiderman 2020).
- LASSO's specific main strength over Ridge regression is that LASSO is able to reduce the number of features (Elumalai 2019).

Disadvantages

- Increased bias over linear regression.
- Worse performance than Ridge regression.
- While linear regression tends to underfit data, polynomial regression has the tendency to overfit the data when the dimension of the regression line increases. This means that the equation has a low bias, fitting the problem's data well, but has high variance as the equation will only fit the current training set well. When data is implemented from outside the training set, for example a new testing (data)set, it would most likely not lay on the overfitting equation.

3.4.4. Polynomial Regression

Polynomial regression builds upon the same principle of multivariate linear regression, but the data points are fitted with a polynomial instead of fitting the data with a linear equation. Polynomial regression is preferred over linear regression when the data's relationship with the dependent variable is too complex for a linear equation. An exemplary polynomial regression algorithm is provided in equation 6.

$$\hat{y} = \alpha_0 + \sum_{i=1}^{m} \alpha_i x_i + \varepsilon_i + \sum_{i=m+1}^{k} \alpha_j f_{P_j}(x_1, x_2, ..., x_m) \tag{6}$$

The first three terms can be recognised from equation 4. The 4th term denotes the addition of polynomial functions to the linear combination while f_{P_j} denotes a polynomial function (Bonaccorso 2017).

Polynomial regression may offer a solution in cases where linear regression underfits the data and therefore is unable to capture the different patterns of the data. More accurate fits can be made by adding polynomials to the equation, increasing the complexity of the model in the process. An example of a polynomial fit can be found in figure 7.

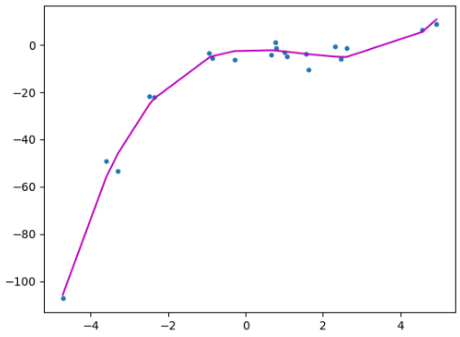


Figure 7 Polynomial Regression Fit, Source: (Agarwal, 2018)

Advantages

- The strength of polynomial regression is that it can fit more complex data relationships better than its linear counterpart.

Disadvantages

- The downside is that polynomial regression is harder to understand and more time-consuming (Elite Data Science 2021).
- While linear regression tends to underfit data, polynomial regression tends to overfit the data when the dimension of the regression line increases. This means that the equation has a low bias, fitting the problem's data well, but has high variance as the equation will only fit the current training set well. When data is implemented from outside the training set, for example a new testing (data)set, it would most likely not lay on the overfitting equation.

3.4.5. Support Vector Regression (SVR)

Support Vector (Machine) Regression (SVR) is derived from its classifier counterpart, the SVM. SVMs are popular algorithms in classification problems that are designed to classify classes based on an optimal hyperplane (Cortes and Vapnik 1995). An impression of such a hyperplane can be found in figure 8:

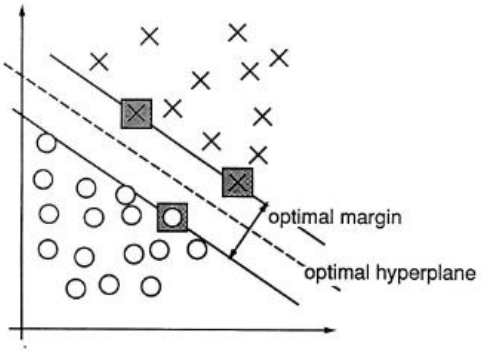


Figure 8 SVM Hyperplane, Source: (Cortes and Vapnik 1995)

While the classification model returns a value from a finite set, namely one of the several classes addressed in the problem, its regression algorithm aims to predict a "continuous-valued multivariate function" (Awad and Khanna 2015, p.67). In SVM, the algorithm aims to find the hyperplane that can distinguish the most amount of data points correctly in the training set. SVR also aims to create this hyperplane but aims to fit a hyperplane through the "maximum number of points" (Raj 2020a).

Besides the input data, the algorithm makes use of a set of hyperparameters that may be tuned, as finding the optimal configuration may improve the model's performance significantly. SVR's Python documentation in SciKit-Learn makes use of various hyperparameters ("SciKit-Learn 1.0.1 | Sklearn.Svm.SVR" 2021):

1. C: A regularization penalty, like in LASSO and Ridge regression.
2. Gamma (for 'rbf', sigmoid and precomputed kernels): indicates the amount of spread or curvature of the boundaries.
3. Kernel: Various kernels, types of functions, may be used depending on the type of problem. A linear kernel may be selected when a linear relationship may be identified between input and output. Popular options are linear, polynomial, 'rbf', sigmoid and precomputed.
4. ε: Distance ε around the hyperplane make up the margin between the datapoints.

Advantages

- SVR is expected to perform better in cases with many input features, or high dimensional space (Drucker et al. 1996).
- Provides high prediction accuracy, while being implemented rather easily (Raj 2020b).
- Flexible in a variety of problem cases given its various kernels.
- Less sensitive to outliers compared to linear or polynomial regression (Raj 2020b).

Disadvantages

- SVR tends to be less accurate given large datasets, or when the data is rather noisy (Raj 2020b).
- The algorithm underperforms when the amount of features exceeds the size of the dataset (Raj 2020b).

3.4.6. Decision Tree Regressor (DTR)

Decision trees have been applied in various disciplines. Examples of such applications are risk management, finance, option pricing, and many more. Similar to its more traditional counterpart, Decision Tree Regressor (DTR) makes use of branches and 'Interior nodes' where classically a decision was made. At the top of the tree resides the 'root node', which resembles the entire dataset (Awad and Khanna 2015, p.15). Each node is split into two new nodes, either a new interior node or a leaf node. The leaf node resembles the "average of the value of the dependent variable in that particular leaf node" (Gurucharan 2020). An exemplary DTR algorithm is illustrated in figure 9. It should be noted that small sample size is illustrated in the figure, resulting in an MSE=0 for all leaf nodes.

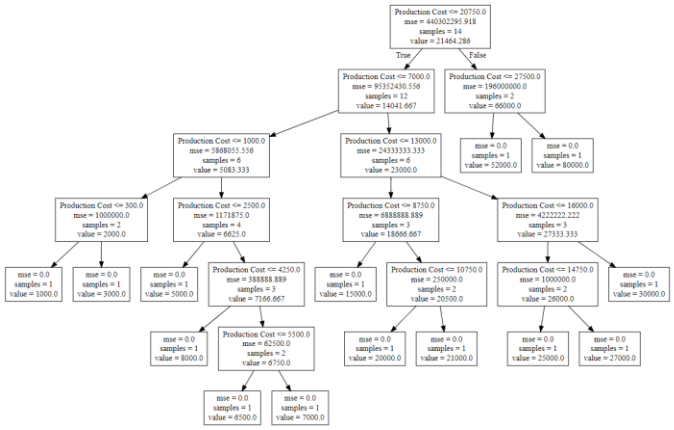


Figure 9 DTR Example, Source: ("Geeks for Geeks | Decision Tree Regression Using SKLearn" 2021)

DTR Python documentation in SciKit-Learn makes use of various hyperparameters("SciKit-Learn 1.0.1 | Sklearn.Tree.DecisionTreeRegressor" 2021). Not all used criteria are equally relevant, the most relevant ones are discussed below:

1. Max_features: The maximum amount of features evaluated at each new split.
2. Max_depth: The maximum depth of the tree, therefore an indication of the number of splits.
3. Min_samples_leaf: Minimum sample size for each leaf node.
4. Max_leaf_nodes: The maximum amount of leaf nodes.

Advantages

- Easy to understand and less data cleaning required (Gurucharan 2020). Given that the decisions of the algorithm are visualised clearly, makes it easy to be interpreted by non-specialists (Lewis, Ph, and Street 2000) (Morgan 2014).
- Almost no hyper-parameter tuning is required, and able to solve non-linear problems (Gurucharan 2020).
- Requires less effort for data preparation (Dhiraj 2019).

Disadvantages

- Tends to overfit data (Gurucharan 2020).
- Predictions of individual DTRs have relatively high variance and bias (Awad and Khanna 2015a).
- Less accurate as regressor compared to other regression algorithms (Morgan 2014) (Dhiraj 2019).

3.4.7. Random Forest Regression

An RFR is a collection of relatively shallow, uncorrelated DTRs (Awad and Khanna 2015a). By averaging the set of uncorrelated DTRs the amount of variance compared to a single DTR is reduced. This process is illustrated in figure 10 below.

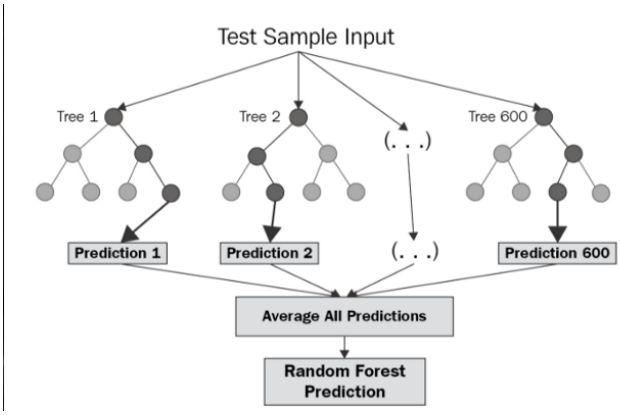


Figure 10 Random Forest Consisting of 600 DTRs, Source: (Bakshi 2020)

Whereas a single DTR is highly interpretable as a result of the clear visualization of the algorithm's workings, this is not the case for RFRs. An RFR consisting of n-trees has n-visualizations. As a trade-off, it does increase the performance of the estimator.

RFR's Python documentation in SciKit-Learn makes use of hyperparameters similar to DTRs ("SciKit-Learn 1.0.1 | Sklearn.Tree.DecisionTreeRegressor" 2021).

Advantages

- RFR are better predictors than DTRs as they are less likely to overfit (Breiman 2019). This is a result of averaging a set of uncorrelated DTRs.
- Similarly to DTRs, less effort is required for the pre-processing of data. This is a result of DTRs, and Random Forests subsequently, being able to work around missing values.
- Random Forests can perform a form of feature prioritization, outputting the relative importance of each input feature for the output variable (Dhiraj 2020a).

Disadvantages

- Random Forests are better classifiers than regressors (Breiman 2019).
- Random Forests, both classifiers and regressors, become rather complex (Dhiraj 2020b). Consequently, Random Forests also take up more training time.
- Less interpretable than DTRs.

3.5. Conclusion

A broad variety of topics has been discussed in the literature study. Each of the subsections contributes to the research study but in different ways.

The purpose of section 3.2 was to discover what tender price influencing features have been identified by previous researchers. Over 77 features have been identified by the cited authors. An overview of the extensive list of features can be found in Appendix E.1 due to the size of the table.

Section 3.3 illustrates what ML models have been developed in construction tender management before. The features used to create the models can be found in Appendix E.2. Two of the ML models in section 3.3. make use of regression algorithms to predict the price of construction projects. The models of Elhag and Boussabaine can predict the construction prices of schools with corresponding accuracies of 79.3 % and 82.2 %. The model of Matel et al. can predict the costs of engineering consultancy services with a mean absolute error of 13.65 %, i.e. an accuracy of 86.35%.

Section 3.4 provides an overview of the most popular regression techniques in ML. A shortlist of regressions algorithms can be found below:

1) Linear Regression
2) Ridge Regression
3) Regression
4) Polynomial Regression
5) Support Vector Regression
6) Decision Tree Regression
7) Random Forest Regression

4. Selecting Tender Price Features for Model Input

Appropriate tender features are selected for the ML model input in this chapter. The set of tender features as found in literature and illustrated in Appendix E is used as a starting point. However, not all features can be taken into account as input for the ML model. Given that the used tender database is rather small (<1000 projects), it is necessary to filter the set of tender price features even more through expert interviews and check whether the features are suitable to be used as input.

- Section 4.1 describes what requirements are considered to filter the features.
- Section 4.2 provides an overview of the most relevant tender price features according to Dutch tender practice.
- Section 4.3 contains a synthesis of all features found in literature or found during the first round of interviews, complying with the selection requirements.

4.1. Feature Requirements

This section is devoted to the construction of feature requirements. The requirements are used to determine whether features are suitable to use as input for the model. These requirements are applied to the features found both in literature and during the first round of interviews with Dutch Tender practice.

An overview of the set of requirements used to filter the features can be found in table 1.

Table 1 Feature Selection Requirements

Requirement No.	Requirement description	Explanation
1	The feature is mentioned at least twice in literature or during the interviews.	By requiring the feature to be mentioned at least four times in either literature or during the interviews, it is guaranteed that multiple authors or tender professionals share the viewpoint.
2	The selected feature is generic and applicable to all sectors within the civil infrastructure industry.	Data on civil engineering tenders is collected from the database. The choice is made to include tenders of all sectors. Object-specific features may not be represented in other civil engineering sectors. For example, the length of the rail track is irrelevant for a real estate tender.
3	Numerical or categorical data on the feature is present in the tender database	If there is not suitable data present in the tender database, it is not possible to include the corresponding feature as model input.

It is verified in section 4.3 whether features comply with the requirements of table 1. If features comply with requirements 1&2, it is checked whether the features are present in the database of tenders.

4.2. Tender Price Features in Dutch Tender Practice

Besides the tender features identified in the literature study of section 3.2, expert sessions are held to check what different features are most dominant in the current Dutch infrastructure tender market. This is done to complement the identified tender features, all originating from papers published between 1998 and 2005 in foreign markets.

The objective of the interviews is to become more familiar with the Dutch tender practices and investigate which input features are most important to use for the ML model. Besides investigating which input features are most important, questions are asked with regards to the most desirable output of the model.

A consent form is sent to the individual interviewees before the interviews. The form provides information on the content of the interview and how the privacy of the interviewees is safeguarded during the process of the graduation research. An exemplary consent form can be found in Appendix C. The consent form and the data gathering process is conform to the Human Research Ethics Committee (HREC) regarding the storage of interview data.

A template for the (Dutch) interviews is provided in Appendix E. As stated before, the main objective of the interviews is to discover which tender features are most important to predict the tender's price, what type of output is most desirable and where in the tender process the Machine Model may be most beneficial to the contractor's organization. A selection of the most important questions asked to progress is provided below:

1) 'Where in the tender process do you see an opportunity for ML?'
2) 'Currently, the output variable of the ML model is the tender price in €. Would you prefer a different output variable?'
3) 'Could you provide a selection of the most influential tender price factors?'

Six interviews were held with tender professionals from the Dutch contractor with various roles to become acquainted with the tenders in the Dutch construction market. The purpose of the varying roles is to collect insights from a broad set of tender-related disciplines. The interviewees had the following roles:

Table 2 Roles Interviewees

Tender Professional (TP) No.	Role	Experience
TP 1	Management Trainee	0-2 years
TP 2	Management Trainee	0-2 years
TP 3	Design Manager (Tenders)	20+ years
TP 4	Tender Manager (Rail)	20+ years
TP 5	Risk Manager (Tenders)	10-20 years
TP 6	Tender Board	20+ years

The tender professionals are denoted by TP1-TP6 (Tender Professional 1 – Tender Professional 6) to ensure the privacy of the interviewees. Features that have been mentioned in the interview with the two management trainees are considered as a single mention. Transcriptions of the interviews with the contractor's tender professionals are included in Appendix D (**CONFIDENTIAL)**. The transcriptions are not included in the public thesis repository.

It was mentioned by some tender professionals that the ratio between the winning tender price and the contractor's tender price could be interesting to predict as an output variable of the model. This was attempted in parallel with the development of the tender price models, but the designed 'tender price ratio'-model was not able to make accurate predictions (R-Squared < 0) with regards to this variable. A negative value for the R-Squared metric implies that the average value is a better prediction than the predictions provided by the model.

4.3. Final Selection of Features

With the most relevant tender price features according to Dutch tender professionals found, it is possible to combine the findings into a final set of tender features. Within this final set of tender features, three different origins are taken into account. First of all, tender features from traditional price estimating research are considered (Appendix E.1). Second of all, features used as input of previous ML models (Appendix E.2). Third, the tender price features according to Dutch tender practice are used (Appendix E.3).

It is verified whether the tender price features comply with the suitability requirements of table 1 to assess which tender price features are suitable to be used as input. Features are considered as input only if the variable complies with all three requirements. If this is not the case, the feature is dropped from the final set of features.

An overview of the final set of features, including the scores on the requirements and a description of the available data, is provided in table 3. An assessment of all features mentioned in Appendix G can be found in Appendix F.

Table 3 Overview Selected Features

Feature No.	Feature name	No. of occurrences	Generic feature	Data type	Data type available
6	Complexity	8	Yes	Categorical	• 'Tender Category'
9	Form of procurement	8	Yes	Categorical	• EMVI or Price-only
11	Project size	8	Yes	Numerical	• Ballpark estimate (€)
12	Duration	9	Yes	Numerical	• Duration (months)
14	Type of object	9	Yes	Categorical	• Type of object
17	Type of client	4	Yes	Categorical	• Type of client
4	Number of tenderers	5	Yes	Numerical	• Total No. of tenderers (n)
63	Type of contract	7	Yes	Categorical	• Type of contract • Contract scope

For all features, except 'Complexity', the available data is rather straightforward. Each of these features is represented explicitly within the tender database provided by the contractor. The feature 'Complexity', however, is rather ambiguous. Complexity can be interpreted in various ways. As a result, the so-called 'Tender Category' of the corresponding projects is taken into account.

The 'Tender Category' is a qualitative label given to categorise potential projects. Tenders may be given 5 different types of labels, from 'Category E' up until 'Category A'. Category A is given to the largest, most complex tenders while Category E is given to small, non-complex tenders. The following criteria are taken into account:

- Order Value
- Contract Type
- Region
- Logistics
- Ground conditions
- Client Track Record

- Contract Experience
- Risks
- Organization Complexity
- Technological Complexity
- Client Relationship

The variable Tender Category is taken into account as a degree of total project complexity, as the separate scores on the criteria are not provided in the database. Ideally, technological/organizational are taken into account as stand-alone features but these are not found explicitly in the database.

4.4. Conclusion

With the results of chapter 4, it is possible to answer the first subquestion:

1) What tender price features influence the tender's price?

Appropriate tender features have been discussed in both literature and within the interviews. The main addition of the interviews is to get an insight into the Dutch tender practices. Taking into account the availability of the data, whether the features are generic and the number of occurrences in both literature and interviews and quantifiability of data, the tender features of table 8 have been selected to use as input for the ML model.

Table 4 Final Set of Tender Features

Feature name
Complexity
Form of procurement
Project size
Duration
Type of object
Type of client
Number of tenderers
Type of contract

5. Preparation Model Development

This chapter is devoted to the preparation of the ML development. The main purpose of the coming sections is to make sure both the data and the models are prepared in the right manner to start predicting tender prices.

- Section 5.1 is devoted to the construction of the requirements of the ML model.
- Section 5.2 introduces the data preparation steps which need to be completed in order to obtain a dataset useable as model input.
- Section 5.3 investigates the quality and the size of the tender dataset used as input for the model.
- Section 5.4 contains a comparison of the possible algorithms, with the purpose to select the three most suitable algorithms.
- Section 5.5 discusses the findings.
- Section 5.6 presents the conclusion.

5.1. Requirements ML Model

The requirements of the ML model are based on the needs of the user, as explained in chapter 2. The needs of the users are collected during the exploratory interviews with the tender professionals to determine what they would like to see in an ML model supporting their daily practices.

An overview of the collected needs can be found in table 5:

Table 5 Overview Users' Needs

Need No.	Description
1	The ML model should be able to predict tender prices accurately.
2	EMVI-tenders should be discounted by their fictional discounts.
3	The ML model should be designed to be implemented in an early stage of the tender process, to support the decision to tender.

The identified needs of table 5 are transformed into measurable requirements to evaluate the model's performance in table 6.

Table 6 Requirements ML Model

Requirement No.	Description	Test Method	Source
1	The model can predict the tender price with an accuracy of at least 70%.	Comparing the model's predictions with the actual tender prices.	Appendix H
2	All EMVI-tenders are discounted by their fictional discounts.	Check whether all EMVI-tenders are fictionally discounted.	Appendix D
3	The input data of the ML model is available before the decision to tender.	Check whether all input data is available before the decision to tender.	Appendix D

After the development of the ML model, it is verified whether the Tender Price Predictor complies with the requirements of table 6.

5.2. Data Preparation

The data preparation phase entails preparing the collected data in useable input for the algorithm. A set of activities is conducted to achieve this (Brownlee 2020c, p.17):

1. Cleaning the Data
2. Feature Selection
3. Data Transformation
4. Feature Engineering
5. Dimensionality Reduction

Data Cleaning

The raw data initially derived from the database may be polluted; i.e. may consist of duplicate rows, wrong entries and errors. The 'Data Cleaning' step consists of investigating and modifying the data in such a way that clean data remains. As a result, the final dataset is much smaller in size and dimension than the original 'raw' dataset.

The following data cleaning steps are considered (Brownlee 2020c, p.18):

1. Identifying outliers through statistics
2. Removing duplicate columns
3. Removing duplicate data points or rows
4. Marking missing values
5. Replacing or deleting missing values

A rule of thumb for the amount of data required in regression analyses is as follows: per predictor/type of variable, 10 occurrences should be in the dataset (Mitsa 2019).

After the raw data has been cleaned, features are selected.

Feature Selection

Feature selection entails decreasing the number of dimensions of the data matrix. Dimension reduction implies that unnecessary features, which may strongly correlate with other features or weakly influence the tender's price, are omitted from the dataset to reduce the computational costs. Feature selection can be done manually, based on the underlying relationships of input features with the output variable. This feature selection method is called 'feature filtering' (Brownlee 2020c). Feature filtering results in a smaller subset of more relevant features for the output variable.

Data Transformation

Data transformation entails changing the form of the input feature or output variable. Two forms of data are desired, either categorical or numerical values, for the use of ML modelling. The raw input variables need to be transformed as such that either categorical values or numerical values are obtained.

Numerical values

Numerical values may be obtained by changing the data type of numerical input to integers or floats. When transformed, the numerical values may be 'scaled' to increase the performance of the ML model. The two most popular types of scaling are 'normalizing' and 'standardizing'

Normalizing entails the rescaling of a numerical feature to a range between 0 and 1 (Brownlee 2020c, p.215):

$$y = \frac{x - Min}{Max - Min} \tag{7}$$

With y being the newly normalized value of x, given Min and Max being the lowest and highest values of the numerical feature.

Standardizing entails the rescaling of data by using the statistical properties of the numerical column. The use-case of either standardizing or normalizing depends on the distribution of values. Standardizing is applicable when the data tends to fit a normal distribution:

$$y = \frac{x - \mu_x}{\sigma_x} \tag{8}$$

With μ_x and σ_x being the mean and standard deviation of x respectively.

Categorical values

To assess categorical features in a regression algorithm, the textual inputs should be transformed into numerical values. The three main options to translate categorical values to numerical inputs are (Brownlee 2020c, p.21):

1. Encoding categorical features as ordinal integer feature
2. One-Hot-Encoding (OHE)

Ordinal Transformation

Ordinal transformation concerns the transformation of a categorical feature into an ordinal integer feature. An example of such transformation is the following:

Table 7 Example Ordinal Transformation

Feature name	Column Name	Range values
Temperature (Categorical)	Temperature	Freezing, cold, neutral, warm, hot
Temperature (Transformed)	Temperature	1, 2, 3, 4, 5,

Ordinal transformation mainly works when order is present within the feature. Examples are temperatures, seniority within jobs, income, education, etc. Ordinal transformation doesn't work when there is no clear order for example colour or, more appropriate for this study, types of contract or types of client. For such features, OHE provides a solution.

OHE

OHE makes it able to transform non-ordinal categorical features into binary entries. A categorical feature (single column), with n-unique entries, is transformed into n-1 columns with binary entries. An example is provided in table 8.

Table 8 OHE-Example

Feature name	Column Name	Range values
Type of Contract (Categorical)	Contract	UAV, RAW, UAV-GC, Other, Bespoke_(Custom)
Type of Contract (Categorical)	Contract_UAV, (…) Contract_Other	0, 1

When the contract of use is equal to 'UAV', the transformed feature will score 0 in all 'Contract_' columns except for a 1 in the 'Contract_UAV' column. OHE will increase the dimensionality of the dataset, but does make it able to incorporate non-ordinal categorical features.

Feature Engineering

Feature engineering entails the modification of input features into new input features by means of mathematical operations. This may be done based on the context of certain variables, for example raising certain variables to the power of n or deriving the distance between two date-time inputs.

Dimensionality Reduction

Dimensionality reduction entails reducing the number of features (columns) of the dataset, without losing any information. This may be useful when some features are highly correlated, reducing the complexity of the matrix by the removal of linear independencies (Brownlee 2020b, p.23).

Feature selection can also be automated based on the predictive model's performance. A popular example of such a method of feature optimization is 'Recursive Feature Elimination' (RFE) (Vickery 2020). RFE is a feature selection method that gradually decreases the set of features, optimizing the subset of features for which the prediction accuracy is maximised. The procedure of RFE is conceptualised in figure 11 (Chen et al. 2018).

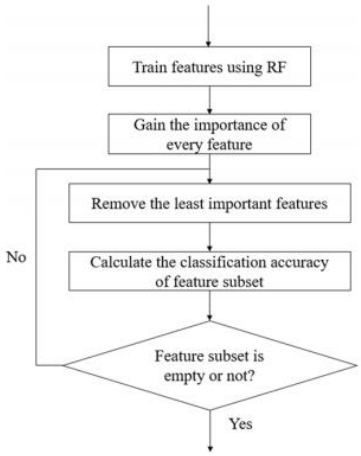


Figure 11 Conceptualised Procedure of Recursive Feature Elimination, Source: (Chen et al., 2018)

Another form of dimensionality reduction is 'Principle Component Analysis' (PCA). The purpose of PCA is to reduce the dimensionality of the dataset by combining some non-correlating features into a linear combination of the original features (Awad and Khanna 2015b, p.31). The result is a so-called principle component (PC), with the original features being omitted from the dataset. This reduces the amount of noise and improves the computational speed of the model.

When the data has been pre-processed, it is possible to initiate the model training phase. The prepared dataset may now be used for the training and testing of the ML model. Of course, this may not be successful in a single iteration. Multiple iterations between training, testing, evaluating and back to data preparation may be required before a final model is obtained.

5.2. Tender Dataset

This section provides more information on the dataset on tenders used as input for the ML models. Ensuring the quality of the data is very important, as 'polluted data' results in worse predictions by the model.

- Subsection 5.2.1 provides background on the used data.
- Subsection 5.2.2 investigates the quality of the data and the causes of omitted tenders.
- Subsection 5.2.3 describes the final tender dataset.

5.2.1. Background on Used Data

Data on the selected tender features have been retrieved from the databases of the Dutch contractor. Within the database, information has been collected on the entire lifecycle of tenders since 2017. Tenders that the contractor did not participate in are tracked as well. The raw data consists of manual inputs, which can either be selected from a drop-down menu by the user or be documented freely (qualitatively or quantitatively) by the user. The inputs are provided by tender managers or tender project team members working on the project. Currently, the purpose of the database is to aid the tender process by keeping track of progress on tenders and their performance in order to improve the decision-making of future tenders.

5.2.2. Data Quality

With the use of Python's 'Pandas' library, data is extracted from the exports of the database. The exports have been provided after consulting an export responsible for the back-end system, taking into account the desired features. The resulting raw dataset consisted out of 12.000 data points on the performances of the contractor and competitors on over 2.700 tenders over the years. In order to convert the raw data into a useable dataset, the data-cleaning steps described in section 5.1 have been completed. The data-cleaning process can be found in Appendix G.

It appeared that the majority of the dataset is unusable as input for the ML model as a result of the cleaning of the data. Figure 12 illustrates that almost 90% per cent of the tender of the initial dataset have been dropped as a result of the data-cleaning process. Upon closer investigation, it appears that the main cause of the invalid data entries is caused by mistakes in the input. Data that has been dropped from the initial dataset has been logged to measure the exact size of invalid data. Figure 12 provides an overview of the separate causes of tenders being dropped from the dataset.

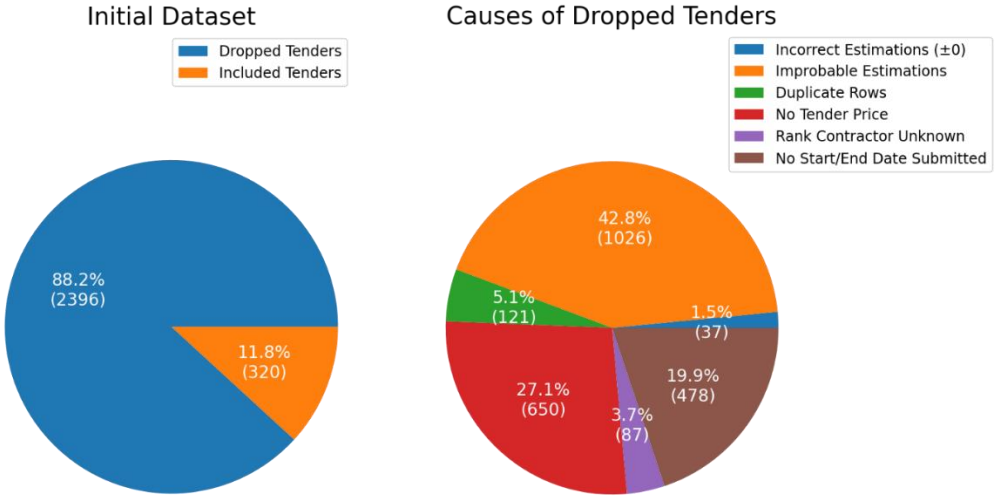


Figure 12 Causes Dropped Tenders, Source: Own Image

A small percentage of the dropped tenders, 1,5%, is caused by extremely low initial estimations of the tender. 37 Tenders have been estimated at prices of 0, 1 or 30 €. These estimated tender prices are extremely unlikely, given that the size of actual tender prices is often multiple millions €.

The majority of the dropped tenders, 42,8%, is caused by unlikely accurate estimations. In these cases, the winning tender price is equal to the estimated tender price. As the estimate in the database is an initial figure to provide a general idea of the size of the project, it is extremely unlikely that this figure is equal to the winning tender price. Including these tenders would have biased the model, by giving the impression that the model is able to provide more accurate predictions than it actually does.

A small part, 5,1%, of the dropped tenders is due to duplicate rows in the database. This could be caused by manual mistakes or duplicate savings of tenders. 27,1% of the dropped tenders is caused by the omission of tenders' prices, while it is known that the tenders are submitted and won or lost. This should mean that a result should be known, but is nevertheless not present within the database. Without a winning tender price, the desired output to predict, it is not possible to take into account these features.

A relatively small part, 3,7%, is dropped as a result of the contractor's ranking of the tender missing. The rank is taken into account to determine whether the contractor has won the tender or not.

Almost 20% of the tender has been dropped due to a missing start or end date of the project. These dates are used to determine the duration of the project. Without one of these two dates, it is not possible to determine the value of this feature.

The resulting 320 tenders are tenders with no missing values for the corresponding features. Tenders with missing values for quantitative features have been omitted from the dataset while missing values for categorical features have been replaced with the value 'Unknown'.

The last applied data preprocessing step which filters the data is detecting and deleting outliers from the dataset, reducing the number of tenders to 222, which make up the input for the model. These steps are described in Appendix G.4. Cases of extreme values for quantitative features or low-frequency values for categorical features negatively influence the accuracy of the predictive model.

5.2.3. Overview Final Dataset

In total, the dataset used as input consists of 222 tenders with 8 features in total, of which 3 numerical features and 5 categorical features. The distributions of the numerical features and the output variable, the tender price, can be found in figure 13.

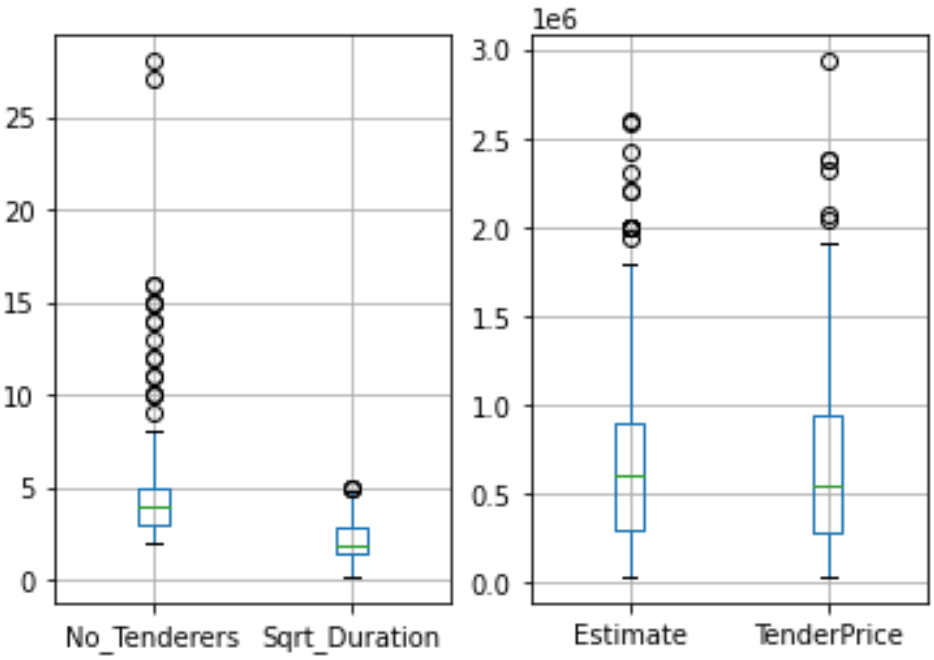


Figure 13 Boxplot Distributions of Numerical Features, Source: Own Image

It should be noted that the feature 'Duration' has been transformed by taking the square root of the initial duration of the project. An explanation for this transformation can be found in Appendix G.4. The categorical features with the corresponding value frequencies are provided in table 9.

Table 9 Overview Categorical Features

Feature Name	Feature's Values	Value Frequencies
Type of Object	Roads	119 (53, %)
	Unknown	103 (46,4 %)
Contract Scope	Build-Only	208 (93,7 %)
	DBFMO	6 (2,7 %)
	Build&DE	5 (2,3 %)
	Build&M	3 (1,4 %)
Type of Contract	UAV	147 (66,2 %)
	RAW	73 (32,9 %)
	Other	2 (0,9 %)
Type of Client	Municipalities	116 (52,3 %)
	Government	64 (28,8 %)
	Province	25 (11,3 %)
	'Construction'	12 (5,4 %)
	Unknown	5 (2,3 %)
Tender Category	Unknown	153 (68,9 %)
	Category D	45 (20,3 %)
	Category E	24 (10,8 %)
Procurement Form	Price-only	176 (79,3 %)
	EMVI	46 (20,7 %)

5.3. Selection Algorithm

Numerous ML models exist, each with its strengths and weaknesses. A preliminary shortlist of popular regression algorithms and their strengths and weaknesses are provided in section 3.4. This list is non-

exhaustive. In general, four selection criteria are used to choose the most suitable ML model (Metwalli 2020):

1. **The data**
   The choice of supervised learning, regression algorithms to be exact, was driven by the type of data and the desired output variable of the tender's price in €.
2. **Required accuracy**
   Not all models are equally accurate, with a trade-off between accuracy and the speed of the model. The accuracy of the models is considered to be one of the most important aspects of the Tender Price Predictor.
3. **The speed of the model**
   The time it takes to make n-predictions is not of great importance, as the model is expected not to run continuously but only to make predictions on demand.
4. **Features and parameters**
   The more parameters are used within a model, the more time is needed to train an ML model. The parameters do improve the flexibility of the corresponding algorithm.

Of these four criteria, just the second criterium 'Accuracy' is considered to be relevant to the problem of this research. Besides these criteria, the aspect 'Interpretability' is taken into account to select the most suitable algorithms. It should be noted that a trade-off exists between these two aspects: the more interpretable a model is, the less accurate its predictions are and vice-versa (Sajee 2020). Table 10 provides a short explanation of the algorithm selection criteria.

Table 10 Algorithm Selection Criteria

Criteria	Explanation
Accuracy	The accuracy is one of the most important aspects of the Tender Price Predictor, as its accuracy should be of an adequate level in order to substantiate tender related decisions within the process.
Interpretability	The interpretability of the results and the model are important to consider, as the user should understand how the model works and how input relates to output.

The advantages and disadvantages of the previously introduced regression algorithms, regarding these criteria, are summarised in table 11 on the next page.

Table 11 Accuracy and Interpretability of Regression Algorithms

Model	Accuracy	Interpretability
Linear Regression	The simplest form of regression, but lacks the ability to make relationships between more complex problems (Elite Data Science 2021).	High interpretability: coefficients of the input features can be acquired to illustrate how these influence the output variable.
Ridge Regression	Ridge regression is able to better fit more complex data than linear regression as a result of its L2-regularization. However, regularization of the penalty function does increase the complexity of the model and increase the bias of the model.	Model interpretability of Ridge is low, as the method does not provide feature optimization.
LASSO Regression	LASSO's L1-regularization decreases the number of features, resulting in a simpler model (Krueger 2021). However, LASSO tends to perform worse than Ridge regression.	LASSO regression is more interpretable than Ridge regression, as the algorithm is able to optimize the subset of features used (Krueger 2021).
Polynomial Regression	Polynomial regression is able to fit more complex data relationships than linear regression but tends to overfit the data at the same time.	Polynomial regression is harder to understand than linear regression (Elite Data Science 2021).
SVR	SVR provides high prediction accuracy while being implemented rather easily (Raj 2020b).	Generally, Support-Vector algorithms tend to be less interpretable than the average regression algorithm. But by using the 'linear'-kernel or 'poly'-kernel, it is possible to derive the coefficients of the features.
DTR	Less accurate as regressor compared to other regression algorithms (Morgan 2014) (Dhiraj 2019)	DTR's decisions are visualized and easy to understand (Morgan 2014) (Gurucharan 2020).
RFR	RFRs are better predictors than DTRs as they are less likely to overfit (Breiman 2019)	RFRs are not easy to interpret, as a RFR consists of a large amount of DTR's.

Both 'Kernel-Based Methods', implying the SVR, and DTRs are close to an optimum of both accuracy and interpretability (Rane 2018). Besides the advantages and disadvantages of algorithms, figure 15 on the next page further illustrates the trade-off between interpretability and accuracy and how the algorithms' performances on these criteria relate.

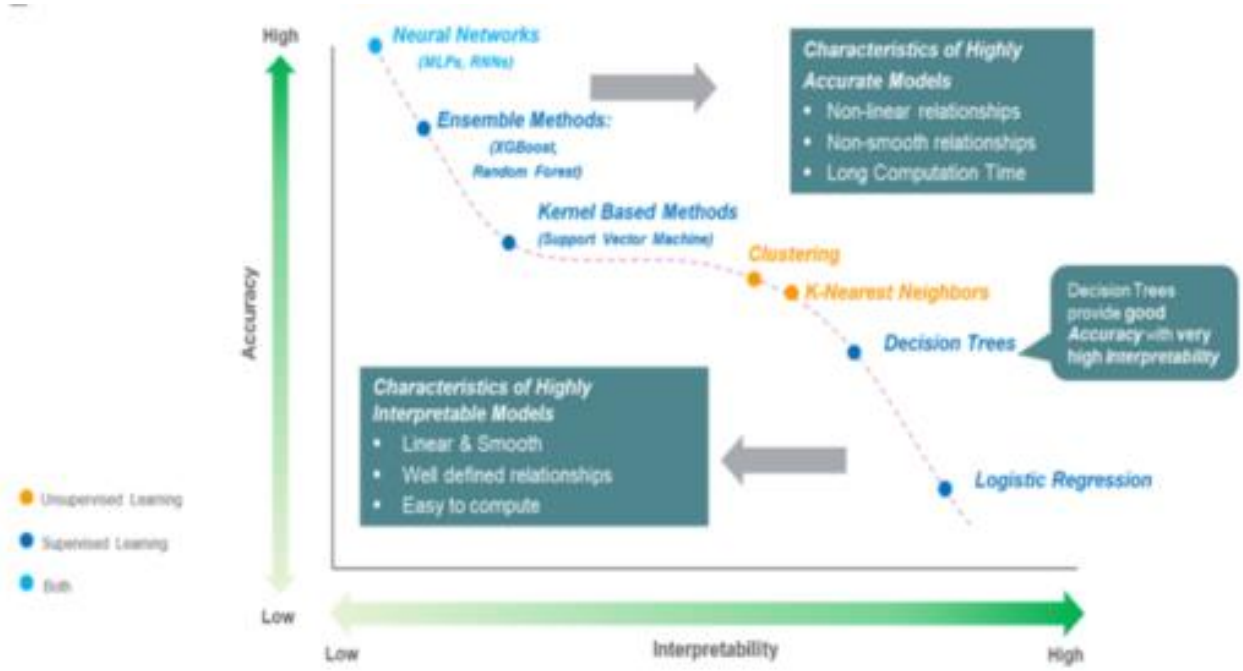


Figure 14 Interpretability – Accuracy Trade-off, Source: (Rane 2018)

Taking into account the criteria of table 8, algorithms are selected to be developed into tender price predicting models. The selected models including substantiations for the selection are illustrated in table 12.

Table 12 Explanation of Algorithm Selection

Model	**Explanation**
Linear Regression	Linear regression is selected over its polynomial or L-regularization counterparts due to lower interpretability and increased complexity of the models.
DTR	DTRs decisions are visualized and easy to understand although worse predictors compared to other regressors (Morgan 2014) (Gurucharan 2020). DTRs are preferred over RFRs as RFRs are hard to interpret (Rane 2018).
SVR	SVRs appear to provide the highest accuracy of all the models while retaining interpretability by selecting its 'linear'-kernel or 'poly'-kernel.

The three selected regression algorithms are developed into tender price predicting models, using the dataset of section 5.2 as input.

5.4. Conclusion

With the results of chapter 5, it is possible to answer the second subquestion:

2) What Machine Learning algorithms are most suitable, taking into account the available data of the contractor?

Three algorithms are used to develop three separate 'Tender Price Predictor' models. Tender data from a Dutch contractor is used as input for the models. Initially, the tender dataset consisted of 2796 individual tenders. After data-cleaning and outlier detection only 222 tenders were deemed as usable input for the ML model, with three numerical features and six categorical features per tender. An overview of the causes of dropped tenders can be found in figure 16 below.

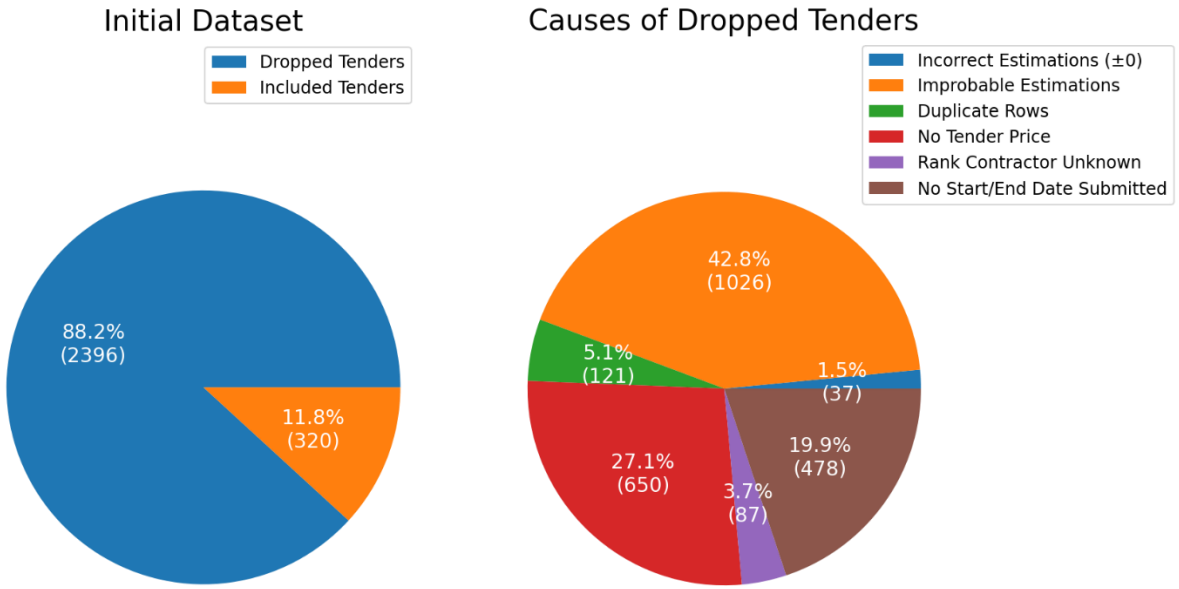


Figure 15 Causes Dropped Tenders, Source: Own Image

The main criteria used to select the algorithms are 'interpretability' and 'accuracy'. The models scoring best on these criteria are Linear Regression, DTR and SVR. Linear Regression and DTR are easy to interpret by deriving the feature coefficients, SVR should provide the highest accuracy of the three according to literature.

6. Development 'Tender Price Predictor' Machine Learning Models

Chapter 6 is devoted to the development of the three regression models to predict the winning tender price. With the dataset completely cleaned and preprocessed it is possible to start developing the ML models to make predictions on the two output variables.

- Section 6.1 describe the ML development steps which are followed in order to design an ML model.
- Section 6.2 provides the results of the three developed ML models.
- Section 6.3 contains an evaluation of the models' performances.
- Section 6.4 compares the model's predictions and the tender price estimations by the contractor's experts.
- Section 6.5 discusses the findings.
- Section 6.6 contains the conclusion.

6.1. Overview Development Steps

Subsections 6.1.1-6.1.5 provide an overview of the separate optimization steps that are followed to develop and optimise the ML Models. The models are optimised to reduce the amount of overfitting while increasing the accuracy of the model.

6.1.1. Model Training and Testing

The 'Model training' part of the development of an ML model revolves around the simulation of the model's experience. Training the model involves entering a large portion of the total dataset, its features and the corresponding values of the output variable, and running the model. The majority of the dataset is used to train the algorithm, with most ML models using between 75% and 80% of the dataset for training purposes (Kamiri and Mariga 2021).

The trained ML algorithm is tested on the remaining 20% - 25% in the model testing phase. Based on the trained ML model, the output variables of the unseen test set are predicted based on the configuration of input features. In the case of small datasets, it is possible to apply the k-fold cross-validation approach to prevent overfitting (Mitchell, 1997, p.111).

6.1.2. Tuning hyperparameters

Hyperparameters are algorithmic parameters that influence how the models learn. The hyperparameters are tuned via the GridSearchCV library ("SciKit-Learn 1.0.1 | Sklearn.Model_selection.GridSearchCV" 2021). Not every algorithm, however, has the same hyperparameters. GridSearchCV finds the optimal combination of the hyperparameters which optimizes the performance of the model when provided with the set of hyperparameters and a provided regression metric. After optimization, the set of hyperparameters is implemented to improve the performance of the model.

6.1.3. Recursive Feature Elimination

Optimal feature selection is achieved through cross-validated RFE, as explained in section 5.1. RFE is performed on both DTR and Linear Regression, but cannot be performed on all SVR-hyperparameter combinations. This is only the case for a 'linear'-kernel, as the hyperplane is constructed in the same dimensional space while this is not the case for the other kernels. Based on the outcome of the hyperparameter tuning, RFE is applied to all models or just DTR and Linear Regression.

6.1.4. K-Fold Cross-Validation

K-fold cross-validation is a way of 'shuffling' the initial test and train sets to get a better indication of how well the model predicts. K-fold cross-validation helps negate the luck effect of random set selection.

Randomly selecting the 'right' test split may provide a biased view of a good fit, whereas selecting the 'wrong' test fold may result in low scores on the evaluation metrics. Cross-validation helps to negate this luck effect by using every train fold also as a test fold (Müller and Guido 2017, p.261). Besides reducing the model's bias, fewer information leaks compared to singular test and train sets as the entire dataset are used for the training and testing of the model.

In k-fold cross-validation, the total dataset is split into 'K-folds' as suggested by its name ("SciKit-Learn 1.0.1 | Cross-Validation" 2021). k-1 splits are used as training sets, while 1 split is used as the test set. A k-amount of training sets results in a k-amount of separate performances, as the test-split rotates for each fold. This process for a 5-fold cross-validation is illustrated in figure 17 ("SciKit-Learn 1.0.1 | Cross-Validation" 2021).

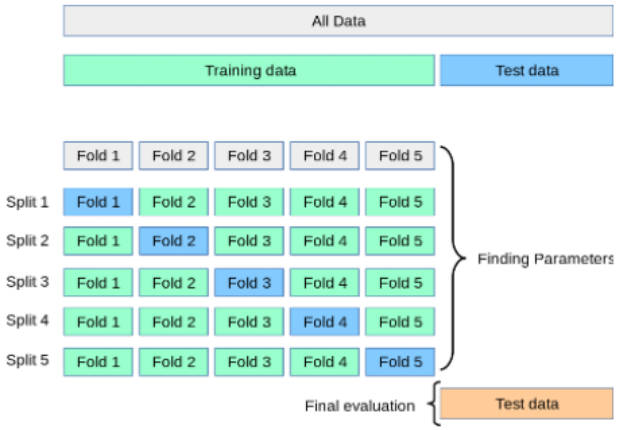


Figure 16 Illustration K-Fold Cross-Validation, Source: ("SciKit-Learn 1.0.1 | Cross-Validation" 2021)

The final evaluation is based on the average performance of the k-folds. In the case of figure 7, the output of the 5-fold cross-validation process is an array of five performance evaluations. This should provide a more robust indication of the model's performance than a single model's testing. The optimal value of k may be determined by means of a sensitivity analysis compared to the average performance of the corresponding k-fold (Brownlee 2020b). This analysis is performed for the hyperparameter-tuned models to inspect how the model's performance is affected by the number of folds.

6.1.5. Model Evaluation

The model's predictive performance is analysed in the model evaluation phase. For regression models it is possible to determine the error or deviation by using regression metrics. In general, three metrics are considered to be the most suitable to evaluate regression models (Wu 2020):

1. Mean Absolute Error (MAE)
2. Mean Squared Error (MSE) / Root Mean Squared Error (RMSE)
3. R-Squared (R^2)

Mean Absolute Error (MAE)

The Mean Absolute Error (MAE) measures the average total error of the model's predictions. The MAE score provides an indication of how large the average error is over an n-amount prediction, taking the absolute value of the error (Wu 2020). The MAE is calculated as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{n}|y_i - \hat{y}| \quad (1)$$

The MAE may be used to compare similar algorithms predicting the same values, given an equal test size. The weight of the error of each prediction is equal, namely 1/N.

Mean Squared Error (MSE) / Root Mean Squared Error (RMSE)

The Mean Squared Error (MSE) / Root Mean Squared Error (RMSE) is similar to the MAE but penalizes errors that are large i.e. huge mispredictions. Instead of taking the absolute value, MSE squares the error of each prediction:

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y})^2 \tag{9}$$

The Root Mean Squared Error, as expected, is the rooted MSE:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y})^2} \tag{10}$$

Both metrics explain something about the averaged error of each prediction but do differ in context. Both RMSE and RME are more sensitive to individual higher errors in predictions, while MAE is not. This means that outliers are penalized more severely by RMSE and RME than by MAE (Akhilendra 2021).

R-Squared

The R-Squared, or R^2 metric, is better known as the Correlation Coefficient. The R^2 score of a model provides a degree of how well the regression model fits the data points by determining how much of the variance of the output variable may be simulated by the ML model. The R^2 score of the model is determined by the following equation:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{11}$$

Where $\hat{y}$ is the predicted output variable, y_i is the actual output variable and $\bar{y}$ is the mean value. If the model perfectly predicts the actual output variable, a R^2 score of 1 is obtained. Usually, the R^2 score is between 0 and 1, with the R^2 score being the degree of output variable variance explained by the model (Wu 2020). It is, however, possible to obtain negative values for the R^2 score. This implies that the denominator is larger than the numerator, i.e. that the mean value of y_i a better predictor is than the predicted value.

The R-squared performance of the different models may be compared to decide which of the models fits the data best. The R^2 score is a rather objective metric, not looking at the context or amount of data, which makes it easier to compare regression models. The R-squared metric is used as the optimization metric, as it can be seen as the most "intuitive metric to evaluate regression models" (Müller and Guido 2017, p.306). The error metrics may be used to compare the separate models with each other on which models manage to obtain the smallest error compared to the actual values.

6.1.6. Results

The predictive performance of the model is illustrated by means of a scatter plot, which visualizes the actual output variable of each data point and the predicted value for the output variable. Besides a visualization, the evaluation metrics are calculated for the hyperparameter tuned, cross-validated ML model and compared to the performance of the baseline model.

6.2. Results 'Tender Price Predictor' Models

This section is devoted to the results of the development of the three ML Models. The features that influence the tender features most, according to the model, are determined and used to optimise the models.

The performance of the models is determined after the k-fold cross-validation of the optimised models, increasing the robustness of the predictor. The results are illustrated in a scatter-plot to illustrate how each separate prediction relates to the actual tender price.

- Subsection 6.2.1 describes the development of the 'Linear Regression' -model.
- Subsection 6.2.2 describes the development of the 'DTR'-model.
- Subsection 6.2.3 describes the development of the 'SVR'-model.

6.2.1. Linear Regression

Optimized Set of Features

Recursive Feature Elimination (RFE), as explained in subsection 6.1.3., is performed to determine the ideal configuration of features. The RFE library of SciKit-Learn is used to optimise the value for the R-Squared evaluation metric. An overview of the results can be found in figure 18.

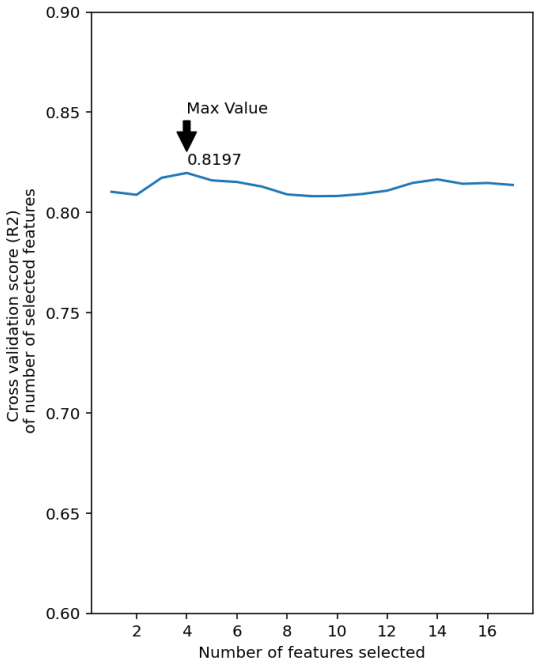


Figure 17 Linear Regression Feature Optimization (Tender Price), Source: Own Image

The maximum R-Squared value of 0.8197, achieved by selecting 4 features, indicates that there is a good fit between the predicted tender price and the actual tender price. An overview of the selected subset of features and their corresponding importance is provided in table 13. The higher the absolute value of the importance, the more impact the specific feature has on the output variable.

Feature Name	Feature Importance
Estimate	0.84
Sqrt_Duration	0.082
Contract_RAW	0.079
Procurement_Price-Only	-0.101

Table 13 Optimal Feature Combination Linear Model (Tender Price)

From table 13 it can be concluded that the initial estimate is considered to be the dominant factor in predicting the tender price, whereas the three other features have an importance of ±0.10. The positive sign implies that a higher value for 'Estimate', 'Sqrt_Duration' and 'Contract_RAW' results in a higher tender price and the 'Procurement_Price-Only' feature implies that price only tenders tend to have lower winning tender prices.

Results

As the estimate is considered to be the most influential feature, a k-fold cross-validation sensitivity analysis is conducted for two different models: for the model with the four selected features, and the model with just the initial estimate as input. The results of this sensitivity analysis are shown in figure 19.

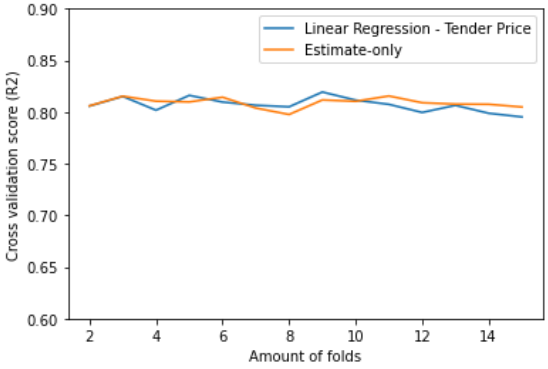


Figure 18 Cross-Validation Sensitivity Analysis (LR) Source: Own Image

The 'Estimate'-only model appears to perform almost as good as the Linear Regression model. The consistency in performance implies that although the folds are reshuffled, in either larger or smaller folds, the model does not appear to be sensitive to changes to its training or testing sets. The scatterplot in figure 20 visualizes how the predicted tender price and the winning tender price relate.

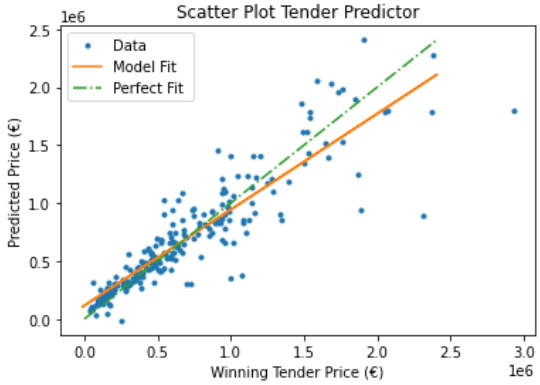


Figure 19 Scatter Plot Predictions Linear Regression, Source: Own Image

6.2.2. DTR

Optimized Set of Features

The result of the feature optimization by RFE has been illustrated in figure 21.

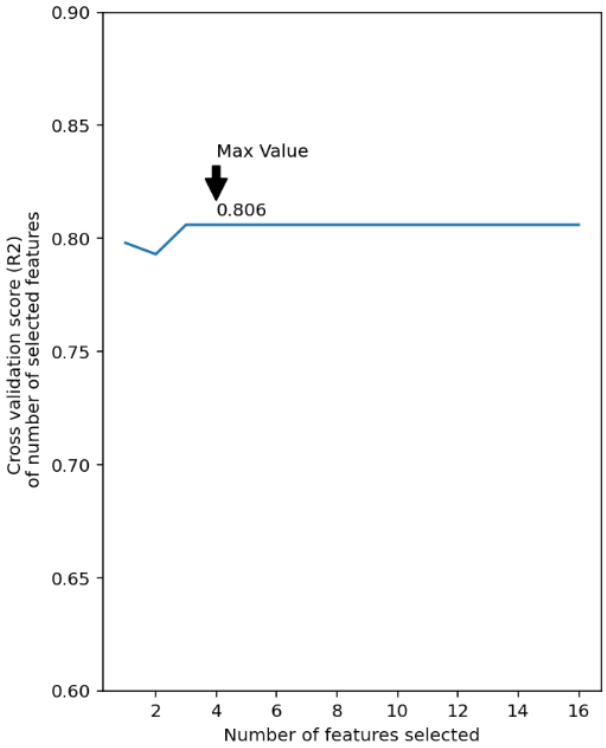


Figure 20 DTR Feature Optimization (Tender Price), Source: Own Image

The maximum R-Squared value of 0.806 is achieved at 4 features, and remains constant after adding more features. An R-Squared value of 0.806 may be considered as a good fit, but worse than the Linear Regression Model. The selected features and their corresponding importance are provided in table 14.

Table 14 Optimal Feature Combination DTR (Tender Price)

Feature Name	Feature Importance
Estimate	0.977
No_Tenderers	0.008
Deliverable_Unknown	0.008
Procurement_Price-Only	0.008

It appears that the tender price predictions of the DTR-model are almost solely based on the 'Estimate' feature, given its importance of 0.977. Compared to the other features, the 'Estimate' feature is rather dominant with importance greater than 100 times.

Results

The dominance of the 'Estimate'-feature is also seen in the k-fold cross-validation sensitivity analysis of figure 22. Again, as the estimate is considered to be the most influential feature, a k-fold cross-validation sensitivity analysis is conducted for both the model with the four selected features and the model with just the initial estimate as input.

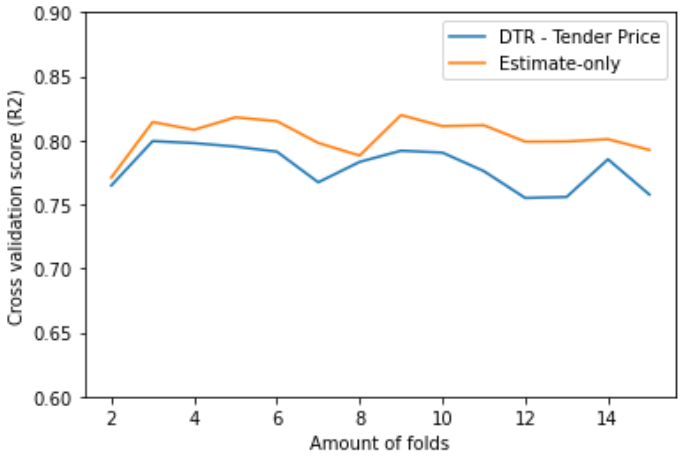


Figure 21 Cross-Validation Sensitivity Analysis (DTR) Source: Own Image

Both models appear to achieve R-Squared scores of ± 0.80, with the 'Estimate'-only model slightly outperforming the DTR model.

The scatterplot in figure 23 visualizes how the predicted tender price and the winning tender price relate.

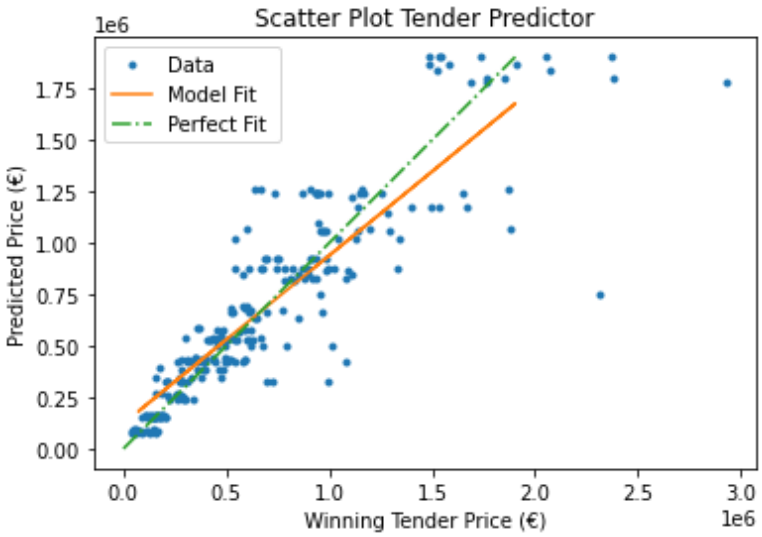


Figure 22 Scatter Plot Predictions DTR, Source: Own Image

Most of the projects appear to centre around the perfect fit line, except for larger tender prices. Horizontal lines can be recognised within the distributions of the data points. This is a consequence of the decision tree method, where multiple projects may end up at the same leaf node resulting in an identical predicted value for all tenders. The decision tree corresponding to the DTR-model is visualised in Appendix I.

6.2.3. SVR

Optimized set of features

The result of the feature optimization by RFE for the SVR-model has been illustrated in figure 24.

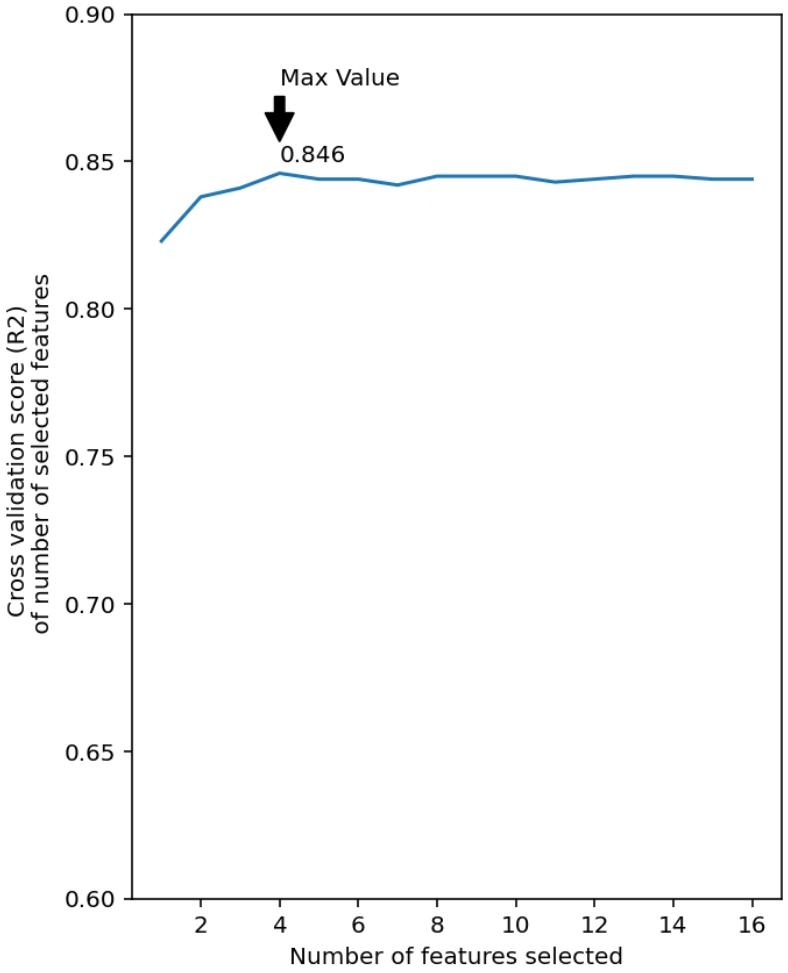


Figure 23 K-Fold Cross-Validation Sensitivity SVR (Tender Price), Source: Own Image

The SVR-model predicting the winning tender price seems to perform better than the previous price predictors. The maximum R-Squared of 0.846 is achieved for a subset of 4 features, illustrated in table 15.

Table 15 Optimal Feature Combination SVR (Tender Price)

Feature Name	Feature Importance
Estimate	0.866
Sqrt_Duration	0.051
Contract_RAW	0.053
Procurement_Price-Only	-0.088

The selected features are identical to the features of the Linear Regression Model, with slightly different feature importance values. Compared to the linear model, the importance values of 'Duration' and 'Contract' are lower while the importance values of the 'Estimate' and 'Procurement' features roughly stay the same.

Results

Just like the previous models, the 'Estimate'-feature is considered to be the most influential feature. a k-fold cross-validation sensitivity analysis is conducted for both the model with the four selected features and the model with just the initial estimate as input.

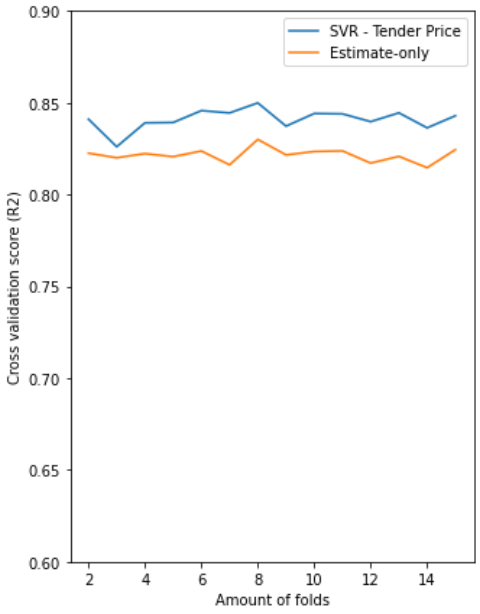


Figure 24 Cross-Validation Sensitivity Analysis (SVR), Source: Own Image

The SVR-model scores better than the 'Estimate'-only predictive model, indicating that the inclusion of the 'Sqrt_Duration' 'Contract_RAW' and 'Procurement_Price-Only' features do improve the accuracy of the predictions.

The scatterplot in figure 26 visualizes how the predicted tender price and the winning tender price relate.

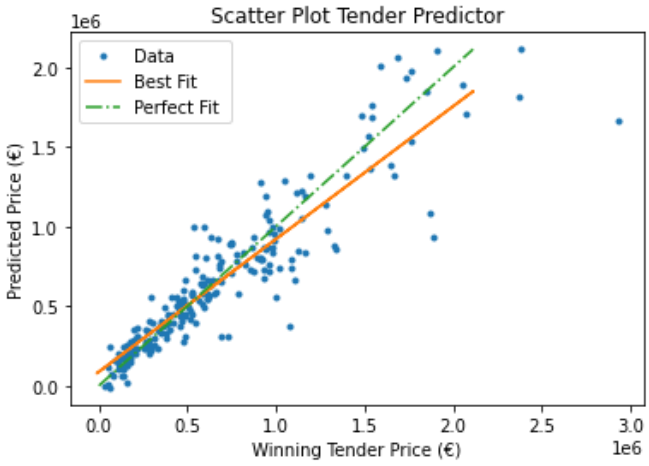


Figure 25 Scatter Plot SVR (Tender Price), Source: Own Image

Similarly to the Linear Model, the SVR-model is capable of fitting the best fit close to the perfect fit. The larger the tenders become, the more the data points seem to differ from the 'perfect fit' line.

6.3. Evaluation of Machine Learning Models

This section contains a final evaluation of the three models, for each of the predicted output variables.

Table 16 Evaluation Tender Price Predictors

Model	R-Squared	MAE	MSE	RMSE
Linear Regression	0.8197	0.265	0.185	0.430
DTR	0.806	0.275	0.202	0.525
SVR	**0.846**	**0.247**	**0.159**	**0.400**

Both the Linear Regressor and the SVR have high R-Squared scores while maintaining low errors. Regarding performance, the SVR algorithm outscores the Linear Regressor slightly on all three regression metrics. The SVR-model is considered to be the model with the highest quality predictions as the model provides the highest goodness-of-fit, given the high R-Squared value, and the lowest errors

6.4. Model Predictions versus Expert Estimations

This section is devoted to a comparison between the model's predictions and the initial estimations of the contractor's experts. Using the cross-validated SVR model, it's possible to predict the tender price for every data point included in the dataset. Given that each datapoint comes with an initial estimate and the winning tender price, it is possible to compare deviation from the tender price estimations and the final tender price. These results can be found in table 17.

Table 17 Comparison Model's Predictions versus Experts' Estimations

	Model's Predictions	Experts' Estimations
MAPE [%]	23,5 %	23,3 %
Std. Abs. Price Error [%]	29,21 %	19,0 %
Min Abs. Price Error [%]	0,390 %	0,742 %
25% Value	8,00 %	8,20 %
50% Value	14,5 %	18,9 %
75% Value	27,6 %	32,4 %
Max Abs. Price Error [%]	308,3 %	97,0 %
Avg. Abs. Price Error [€]	125.581 €	141.964 €
Percentage of Closest Tender Price Estimation [%]	53,6 %	46,4 %

The experts' estimations closely outperform the model's predictions, with a mean absolute percentage error (MAPE) of 23.3% of the winning tender price compared to the model's MAPE of 23.5%. These figures imply that the expert's estimations, on average, are slightly more accurate than the model's predictions. However, when investigating the distribution functions of both tender price approximations, the model's predictions appear to be more accurate than the expert's estimations given that the model P75-value of the SVR model is over 20% smaller than the P75-value of the expert's estimations: 75% of the model's values have a relative error of max 27.6% compared to a max relative error of 32.4% of the expert's estimations.

However, the maximum absolute error of the tender price of the model is much larger than the maximum price error of the expert's estimations. To be precise, one tender is predicted to deviate 308.3 % from the actual winning tender price, compared to a max value of 97.0 % for the expert's estimations. The boxplots of both distributions are provided in figure 27.

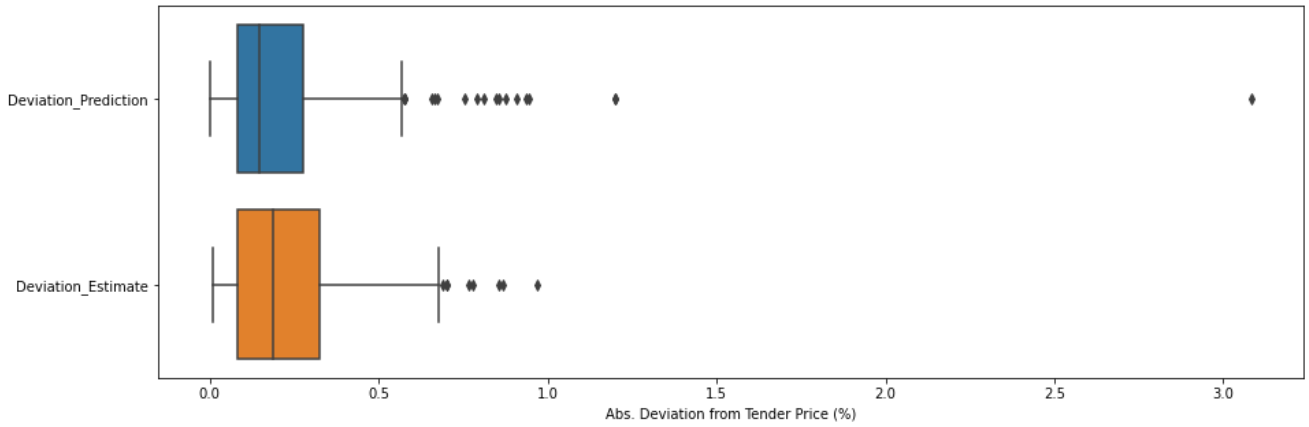


Figure 26 Boxplot Errors Model vs Estimates, Source: Own Image

Apart from the detected outlier, most of the model's predictions are of higher accuracy compared to the experts' estimations. This is can be derived from the lower errors up to the P75 value of the boxplot.

In order to investigate what may have caused to the outlier, details of the project's features have been analysed. The actual tender price is equal to 58.800 €, while the predicted tender price is 240.060 €. The size of the tender is considered to be very small. Suspecting that the size of the tender may have influenced the error, table 18 exhibits the features of tenders of a size smaller than 100.000 €.

Table 18 Outlier Analysis

Tender Price	MAPE	Sqrt_Duration	Contract_RAW	Procurement_Price-Only
53195	93 %	1.71	0	1
35700	94 %	1.42	0	1
58800	**308 %**	**1.42**	**1**	**0**
62000	112 %	0.61	0	1
74800	4.00 %	1.11	0	1
80700	25 %	1.21	0	1
52900	120 %	0.91	1	1
88675	81 %	2	0	1

It appears that the model is bad in predicting tenders of very small sizes given the high values for MAPE. Illustrated in bold, the outlier scores a '1' on 'Contract_RAW' and a '0' on Procurement_Price-only, implying that the tender contains both a RAW-contract while being procured through EMVI which both have positive coefficients resulting in higher tender prices. This explains why the predicted tender price is 308 % higher than the actual tender price. This appears to be a mistake, given that tenders are procured through EMVI only given a large project size. It may be the case that the tender is only a small part of a larger project. Modifying this datapoint improves the model's MAPE from 23.5 % to 22.0 %.

6.5. Conclusion

With the results of chapter 6, it is possible to answer the third subquestion:

3) How accurate are tender price predictions by applying ML algorithms using historical project data?

Three ML models have been developed to predict the tender's price. Of the three models, the SVR-model performed the best with an R-Squared of 0.846. The SVR model includes an optimised set of features, which is a subset of the initial dataset. The model has included 4 features with the following feature importance values:

Table 15 Optimal Feature Combination SVR (Tender Price)

Feature Name	Feature Importance
Estimate	0.866
Sqrt_Duration	0.051
Contract_RAW	0.053
Procurement_Price-Only	-0.088

Noticeably, the 'Estimate'-feature has a larger importance value than the other values. This indicates that the initial estimate influences the tender price more than the duration, type of contract and procurement form.

Comparing the model's predictions to the actual tender price, a mean absolute percentage error (MAPE) of 23.5 % is obtained which is equivalent to an accuracy of 76.5%. This value is marginally lower than the MAPE of the experts' estimations, which obtain a MAPE of 23.3 %. Although the absolute mean deviation of the model's predictions is larger than the experts' estimations, the majority of the model's predictions are more accurate than the expert's estimations (53,6% vs 46,4%). Also, the average absolute price deviation in € of the model is lower than the experts' estimations (125.581 € vs 141.964 €).

7. Evaluation of the Tender Price Predictor

The purpose of this chapter is to validate whether the ML model could be used in Dutch tender practices. Tender professionals of a Dutch contractor have been interviewed to investigate how such models could be used in practice and determine which obstacles still need to be cleared in order to implement the tender price predictor.

- Section 7.1 contains the verification of the ML model requirements.
- Section 7.2 presents the results of the validation interview.
- Section 7.3 concludes the findings.

7.1. Verification ML Model Requirements

The constructed ML model requirements are evaluated in table 19.

Table 19 Verification ML Model Requirements

Requirement No.	Description	Test Method	Test Score	Fulfilled (yes/no)
1	The model can predict the tender price with an accuracy of at least 70%.	Comparing the model's predictions with the actual tender prices.	76.5%	Yes
2	All EMVI-tenders are discounted by their fictional discounts.	Check whether all EMVI-tenders are fictionally discounted.	Fictional discounts of EMVI-tender are included in the tender price.	Yes
3	The input data of the ML model is available before the decision to tender.	Check whether all input data is available before the decision to tender.	All data is available before the decision to tender.	Yes

As the ML model passed all requirements, interviews are conducted to verify whether and how the Tender Predictor may be used in practice.

7.2. Interview Results

The objective of the validation interview is to get an impression of how effective the developed Tender Price Predictor is, and how Dutch tender professionals could apply the tool in practice.

In order to determine how the Tender Price Predictor may effectively be used in Dutch tender practices, a set of questions are asked to some of the contractor's employees which are active in the tender industry. The interviews are planned to take 30 minutes to complete, consisting of a 10-minute introduction to the research and its results so far, with 15-20 minutes room for a discussion and questions. The questions asked to the interviewees are the following:

1) What is the desired accuracy for a usable, predictive model in tender management?
2) What attributes of the Tender Price Predictor do you deem unnecessary, and what has been missed?
3) How can the Tender Price Predictor be used to support the tender professional's decision-making?
4) What are the obstacles for the Tender Price Predictor?
5) How could the quality of the input data be improved?

Noticeably, the first four questions are regarding the potential use of the Tender Price Predictor and its shortcomings. The final question concerns the quality of data as this has proven to be a bottleneck with the majority of the database unusable due to missing or incorrect values.

Five interviews were held to investigate if and how the tender price predictor may be used in practice. Seven professionals were interviewed in five sessions, with interviewees 5-6-7 clustered in the fifth interview. All interviewees either worked on tenders themselves or are directors of regional departments. Selecting not only potential users, the tender managers, but also the management of regional departments of the contractor have been invited to check whether a support base exists to trust on relatively new technological advancements.

The roles of the interviewees are provided in table 20. It should be noted that two interviewees of the first interview round have also been interviewed for the validation round. IV2 and IV4 have been selected given their roles within specific engineering branches in tenders.

Table 20 Roles Interviewees Validation Interview

Interviewee Validation (IV)	**Role**	**Experience**
IV1	Tender Manager	5-10 years
IV2	Tender Manager (Rail)	20+ years
IV3	Regional Director East	10-20 years
IV4	Design Manager (Tenders)	20+ years
IV5	Regional Director East (Tenders)	20+ years
IV6	Regional Director East	20+ years
IV7	Chief Office East (Cost Estimating)	40+

The transcriptions of the interviews can be found in Appendix I (**CONFIDENTIAL)**.

7.2.1. Desired Accuracy of Tender Price Predictor

The interviewees state that the accuracy of the Tender Price Predictor depends on when the tool is supposed to support the decisions of the tender manager. Two main implementation phases are identified: both early in the tender process to substantiate the decision to tender or to support the tender team during the tender itself. Predictions of the tender price predictor in an early stage of the tender process are not required to be as exact as they would be in a later stage.

Regarding estimates of the desired accuracy, a minimum of 30% deviation from the winning tender price is required to seriously weigh in its predictions. Interviewees IV5, IV6 and IV7 state that a deviation of ± 10% from the winning tender price should be the minimum accuracy of the predictor. IV2, IV3 and IV4 state that it is difficult to provide an exact figure, but an accuracy north of 70% should be appropriate in order to be considered in the decision making.

7.2.2. Missing Attributes of the Tender Price Predictor according to Practice

A broad set of features are identified by the interviewees that haven't been included in the Tender Price Predictor. The purpose of identifying what features are missed or could have been improved is to take note for further exploitations of the Tender Price Predictor.

According to the interviewees with a background in writing EMVI-plans, IV1 and IV2, more emphasis should be put on the EMVI-criteria of the tenders. Instead of focusing on whether tenders are procured via the EMVI-procedure, the explicit tender components should be included to identify which tenders the contractor performs best on, and which tenders the contractor doesn't perform well on.

The majority of tenderers agree that the selected features are very generic, and more industry-specific components i.e. the amount of rails or volumes of concrete could improve the accuracy. These remarks may be considered for further expansion of a similar predictor within the appropriate industries.

The complexity feature, assessed by the type of tender category, should be determined more explicitly. Currently, the complex components of the 'tender category' assessment cover a smaller part with criteria changing over time. Taking into account multi-disciplinarity by identifying the amount of collaborating business units may be interesting according to IV4, whereas IV2 states that complexity should be included explicitly as the tender category feature also weighs in a rough estimate of the tender.

One important aspect that hasn't been considered before is the time frame of the tenders. IV6 states that tenders should be considered within a maximum time window of 4 years. Tender results older than 4 years may not be relevant anymore due to a changing market and trends changing over time. This requires the dataset to be refreshed from time to time, with tenders deemed 'too old' to be dropped from the database.

7.2.3. Use-Case of Tender Price Predictor within Tender Management

The interviewees agree that the Tender Price Predictor may be beneficial to the tender professionals' decision-making process but modified to the demands of more specific industries i.e. rail tenders for 'department Rail' or civil tenders for 'department Civil'. More project element specific attributes like type of competitors, amount of concrete, km of road could improve the model's usability in tender practices of the corresponding sectors.

IV1 and IV4 endorse this, adding that the tool may also be used by tender teams to reflect and brainstorm upon. Tender teams could use predictions to substantiate their quality plans to adjust the tender price, focus more on quality plans or stop the tender. Within larger tenders, a better explanation could be provided to the tender desk by providing more objective figures to substantiate choices made by the tender team.

IV5 states that the predictor may be useful to substantiate the decision to tender in an early tender phase. If the predicted tender price seemingly exceeds the ceiling price set by the client, it may be decided to stop with the tender early on. Currently, IV5 and IV7 make similar decisions daily. Objective models like the predictor may be of added value to their decision-making processes. At this moment, this mainly happens on basis of experience and gut feeling.

7.2.4. Tender Price Predictor Implementation Obstacles

The obstacles for the Tender Price Predictor may be divided into two categories, its usability or users and the quality of data.

Besides the accuracy of the tender, the process and its users are key to the success of such predictive models. According to IV2, IV3 and IV4 the human factor of managers providing input to the database and operating the model itself should be weighed in. Users may be sceptical against predictive models states IV3. This is proposed by IV5, who thinks that improving the data and more sector-specific options may result in more accurate predictions. These predictions would co-exist with the work of cost estimators, creating a synergy with human work and artificial predictions.

IV2 and IV4 emphasize that the predictor may not perform accordingly when either too much input is required from the user at the same moment resulting in worse data, or when the purpose of the input data is unclear. Without training, knowing the purpose of the data's quality and the model users may result in less useable data.

7.2.5. Improving Data Quality

IV3 suggests to create control queries that analyse whether data is documented in wrong fields or whether the data is correct. This can be seen as an automated approach to checking the input. Back-end wise, it

could be considered to make certain input fields mandatory to fill in from a database perspective. This could solve the problem of empty fields for certain features.

Front-end wise, the threshold should be minimized to submit the data itself. IV7 supports this by stating that if the tender manager has to fill in 20 fields at the end of every tender he might rush these submissions and therefore affect the data quality. IV4 states that up to a year after the completion of a tender, tender managers are asked to retroactively submit figures on tenders. The quality of the data could be improved by defining what data is required or important, and what the rationale behind the data quality is.

7.3. Conclusion

The results of section 7.2 are used to answer subquestion 4:

4) How can the Tender Price Predictor effectively be used within the tender practices of Dutch contractors?

Generally speaking, the interviewees agree that contractors may benefit from the Tender Price Predictor when it is possible to accurately predict, within a relative maximum deviation of ±30% (subsection 7.2.1), the winning tender price. The use-case of the Tender Price Predictor may be improved when project-specific or sector-specific characteristics would be used in order to meet the demands of a more specialised sector within the contractor's organization (subsection 7.2.3). Also, the EMVI-component and complexity feature could be improved upon by including more quality-plan components and replacing the tender category feature with more specific complexity variables (subsection 7.2.2).

The Tender Price Predictor could encounter obstacles upon implementation. Requiring too much effort from the users may result in worse quality of data. Both the users of the model and the managers submitting tender data should be trained accordingly in order to obtain maximum efficiency (subsection 7.2.4).

8. Discussion

Sections 8.1-8.5 contain the discussion of the findings of this research. Section 8.6 highlights the limitations.

8.1. Desired Output Variable of the Tender Price Predictor

The EMVI component of tenders should be taken into account besides the tender price only. This is a derivative of the Dutch tender market's procurement method, as introduced in section 3.1. The EMVI-orientated answers are not strange, given that some interviewees have experience in the writing of EMVI related quality plans for tenders. The focus of the output variable, the tender's price, shall not be changed as an 'EMVI-predictor' is another model on its own. EMVI is a relatively new procurement procedure as it was introduced in 2014 by the European Parliament. This also explains why the quality plans of this procurement procedure are not mentioned in previous studies on price-influencing factors (Odusami and Onukwube 2008; Elhag and Boussabaine 1998; Elhag, Boussabaine, and Ballal 2005; A. S. Akintoye 1999).

Regarding previous literature, some studies catch on with the decision to predict the decisive tender price. For example, both Zhang et al. (2015) and Matel et al. (2019) modelled these tender prices but for other parties than the contractor (Zhang et al., 2015**.** Zhang et al. designed a model that evaluates the tender bids from the client's perspective. So instead of predicting a tender's price for the contractor, it evaluates whether the received bids fall between or outside a reasonable range for bidding quotation. This study, interestingly, has the same output variable but is designed for the client party. The fact that tender price predictions are sought after for both ends of the tender process confirms that predicting a tender's price, disregarding for which party, may be beneficial to either party. The study of Matel et al. concerns the development of an ANN model to predict the project costs of engineering services of a construction project. The study is similar in its choice of the output variable, but differs in scope: instead of focusing on its application for a contractor's tender phasing, it focuses on the tender estimating activities of engineering consultancy firms. The activities, therefore, differ, with more focus on engineering design and project teamwork than on the actual realisation of the project.

Two other studies, the 'Tender win probability' model of Kultin et al. (2021) and the 'Project success predictor' of Wang et al. (2012), are in line with the objective of this research but predict a different output variable. Both studies aim to predict whether success can be achieved during different phases of the project's lifecycle. In the case of Wang et al. (2012), to predict project success, the differences between the actual and estimated values for cost and scheduling are measured (Wang, Yu, and Chan 2012). These values are, together with scope, part of the iron triangle of project management which indicate project success (Atkinson 1999; Pollack, Helm, and Adler 2018). The selected time and cost performance components by the authors are, to predict project success, logical. Similarities between the discussed output variable lies in the cost or price component of projects. Predicting the planning overrun this early on during the tender may not seem desirable, given that a tender is won on price or EMVI.

8.2. Implementation Phase of the Tender Price Predictor

The implementation of the ML model in the tender process is not discussed explicitly, but the ML models predicting tenders found in the literature point towards the results of the interviews implicitly. The authors of the models used to make predictions in tender management, appear to assume that an early tender stage suits price predictions the most where budgeting and tender selection are key activities.

The tender interviewees unanimously agree that the best fit for ML integration in the tender's processes is during its earliest stage. The purpose of early integration is to support the decision-to-tender, to prevent the waste of money by tendering on projects that the cont has not been successful with. Given that all interviewees are, or have been, active in the tender stages of a project it is not unlikely that they share this opinion. If calculators or engineers of later phases would have been interviewed, they might have suggested a stage where they take part in.

The studies of Matel et al. (2019) and Kultin et al. (2021) concerned tools which are to be used during the early tender stages, confirming this viewpoint. Kultin et al (2021). designed a model that attempts to assess the degree of the prospect of tenders, i.e. the probability of winning a tender (Kultin, Kultin, and Bauer 2021). The model is designed to be used within a pre-project tender stage, similar to the proposition of the interviewees.

The 'Engineering services' costs predictor' of Matel et al. (2019). attempts to predict the engineering costs before an engineering project is started. The engineering service costs are used to provide a quotation for the client. This study has similarities with this graduation thesis, but instead of a contractor's operations focuses on the consultation of engineering firms. The tender implementation phase draws parallels with the early tender phase as stated by the interviewees, but then for the tender phase of engineering services.

The studies of Zhang et al. (2015) and Wang et al. (2012) are not necessarily in line with the findings of the interviews, but do not contradict them at the same time. The developed system by Zhang et al. is implemented after the tender bid. This is a logical follow-up given the system's objective, namely the evaluation of tender bids (Zhang, Luo, and He 2015). The purpose of the model of Wang et al. is to investigate whether the project definition in an early stage of the planning process leads to project success(Wang, Yu, and Chan 2012). This is based on the results of a preliminary literature study which showed that early planning improves profitability and affects cost and scheduling process. The authors attempt to make predictions based on initial cost and schedule estimates in the early stage of the project's operations, while this is outside the scope of the tender phasing as defined earlier in this study.

8.3. Feature Selection Process

Three requirements are set to filter the extensive lists of features, with over 90+ different features mentioned in both literature and interviews. Various features that do strongly influence the tender price are omitted from the final list as a result of these filters. Requirement no.2, 'The selected feature is generic and applicable to all sectors within the civil infrastructure industry', makes it able to compare tenders of different sectors but has the consequence that project-specific variables like material quantities and dimensions of the objects are not useable as input for the model.

Of the selected features, 'Complexity' may be considered an ambiguous feature. In the case of complexity, the 'Tender Category'-variable was selected to represent the project's complexity, as no explicit generic complexity variable was found in the contractor's database. However, it is debatable whether the selected feature is a good placeholder, backed by the model not selecting the feature for the optimal subset of selected features.

8.4. Data Preprocessing and Algorithm Selection

The data-cleaning process significantly decreased the size of the tender dataset. Deleting unusable data from the dataset resulted in 88.2 % of the tenders being left out. Of the remaining 11.8 %, approximately 30 % of the data was omitted due to outlier detection. Outliers on both numerical and categorical features have been omitted, but this is the consequence of a lack of data points representing the corresponding features.

Large parts of both the 'Type of Object' and 'Tender Category' features consist of 'Unknown' values. This affects the performance of the predictive models, as projects with missing values for these features are collected under the umbrella value 'Unknown'. As a result, the corresponding features may be of lesser importance according to the model than they could have been when enough tenders with non-missing values would be present in the dataset.

The algorithm selection is based on the prioritization of two trade-off criteria, interpretability and accuracy, applied on a shortlist of regression algorithms. The shortlist of regression algorithms is non-exhaustive, as other more suitable regression algorithms may have been missed during the initial

exploratory phase. It may also be possible that the most suitable algorithm has not been selected during the process. This may be caused by either misinterpreting information on the algorithms, incorrectly assessing the algorithms' advantages and disadvantages or by the omission of relevant algorithms in the overview of regression models.

8.5. Predictions of 'Tender Price Predictor'-Models

Overall, the 'Estimate'-feature is considered to be the most influential feature of all. In the case of the most accurate model, the initial estimate of the experts is considered to have an impact of 10 times the second-most influential factor. The impact of the feature could be explained by the inclusion of project-specific aspects like materials, risks and other quantifications by the cost experts which are not covered by the generic features as selected in chapter 4. It is unknown what the 'Estimate' exactly contains, only that the value is determined in an early stage of the tender process.

Two features that were expected to impact the tender's price are 'Tender Category', as a factor of complexity, and the 'Type of Object'. However, these features were not represented in the optimised subset of tender features as selected through RFE. It appears that there are two possible causes for this: Firstly, it may be the case that these features simply do not affect the tender's price as much as initially expected. Secondly, both of these categorical features consist of mostly 'Unknown' values due to missing cells in the dataset. These 'Unknown'-values make it hard to determine the exact effect of these features on the tender's price due to highly homogeneous data.

Given the results of the scatterplot in figure 25 on p.42, it appears that the model is less able to correctly predict the tender's price for larger tenders compared to smaller tenders. This is contributed to a smaller representation of these tenders in the dataset. Also, the majority of tenders have been dropped due to outlier detection and the preprocessing of the dataset: projects of enormous sizes or very specific contracts are dropped due to an underrepresentation of these values in the dataset, resulting in even fewer tenders of such sizes.

Previous studies regarding the development of ML within tender management have prioritised accuracy over interpretability. The developed models of Elhag and Boussabaine (Elhag and Boussabaine 1998) and Matel et al. (Matel et al. 2019) make use of ANNs to predict the project's costs. These neural networks have improved accuracy over the selected algorithms but tend to be less interpretable. The ANNs of Elhag and Boussabaine predicting construction costs of schools managed to achieve accuracies of 79.3 % and 82.2 %, while the developed ANN of Matel et al. achieved a MAPE of 13.65 % i.e. accuracy of 86.35 %. These accuracies are higher than the achieved MAPE of the model. This can be explained by the inclusion of more project-specific components, the use of ANNs over interpretable regression algorithms or by a difference in objective. The achieved R-Squared of Matel's model, 0.99, could be caused by the objective of the model, predicting the costs of engineering services of an engineering firm compared to the model's objective of predicting tender prices. An R-Squared of 0.99 is extremely high, implying that 99% of the output variable's variance is explained by the model.

8.6. Limitations

8.6.1. Limitations 'Tender Feature'-Literature

Literature on price influencing factors in construction price estimating has been studied to collect the most influential features. The papers have been published between 1998 and 2005, which means that the sources are between 17-24 old at the time of writing. Certain trends in the construction industry may not be covered, and therefore not be included in the set of tender features. Also, the studied tender price influencing features cover foreign markets like Nigeria and the UK, with no literature on the Dutch market. Specific Dutch social, economic and geolocational aspects are therefore not included in the literature study. As a result, the only Dutch-specific input comes from the interviewees.

Given the small sample size of interviews (n=7), it may be the case that relevant features have not been included in the final list as a result of the mismatch between the Dutch market and foreign markets. To ensure this, extra interviews could have been conducted to include more tender professionals with varying roles and sectors to provide a more holistic view on the most important general tender features. This was not possible due to time constraints.

During the interviews, both tender managers and (local) management were interviewed. This was done to explore the support base for the implementation of predictive models within tender management. Cost estimators or cost engineers, however, have not been interviewed. Their views regarding the most important tender features could have provided new insights regarding the Tender Price Predictor's objective, implementation or different features than previously identified.

8.6.2. Limitations Data Availability

It should be noted that some important features, according to practice and literature, are omitted from the final dataset. Although the final set of features complies with the requirements of generic features and sufficient occurrences in either literature or the interviews, they were not present in the database of the contractor. These omitted features are: 'Project team experience' and 'Location'.

Invalid data and outlier detection resulted in a significant decrease (± -90%) of the useable dataset size. As a result, the ML models have used data on relative small-sized tenders due to large tenders not being represented in the dataset. To incorporate large tenders in the model's predictions, either missing data should be restored accordingly or larger tenders should be divided into smaller modules. Invalid data could have been checked manually by analysing project documents or by interviewing the responsible tender managers, but this has not been attempted due to time constraints.

Interviewees proposed to include more EMVI-criteria in the development of future models. The purpose of adding more EMVI-criteria is to investigate how certain EMVI-components influence the winning tender price. Knowing what tenders suit the contractor best could drive the organization to pursue certain tenders more than others. Currently, the only EMVI-aspects taken into account are the amount of total fictional discount and whether the tender was procured through the EMVI-procedure or not.

8.6.3. Limitations Similar Studies

A lack of similar studies makes it hard to compare the results of the models. The developed models of Elhag and Boussabaine (Elhag and Boussabaine 1998) and Matel et al. (Matel et al. 2019) make use of ANNs to predict the project's costs, but do not focus on the contractor's perspective of infrastructure projects. The performances of both models have been compared by the findings of this study, but the different scopes make it hard to well-founded conclusions.

9. Conclusion & Recommendations

This chapter concludes the research's findings and provides recommendations for further use. Recommendations for both the Dutch contractor as well as recommendations for further research are provided in this section.

- Section 9.1 answers the research question.
- Section 9.2 contains recommendations for the Dutch contractor.
- Section 9.3 provides recommendations for further research.
- Section 9.4 contains a reflection on the research process.

9.1. Conclusion

This section will answer the research question:

How can a Machine Learning algorithm, predicting the tender's price using tender project data, be developed to support the contractor's decision to tender?

Four subquestions are answered in order to provide a conclusion on the main research question. The following subquestions have been constructed:

1) What tender price features influence the tender's price?
2) What Machine Learning algorithms are most suitable, taking into account the available data of the contractor?
3) How accurate are tender price predictions by applying Machine Learning algorithms using historical project data?
4) How can the Tender Price Predictor effectively be used within the tender practices of Dutch contractors?

9.1.1. SQ1: What Tender Price Features Influence the Tender's Price?

Appropriate tender features have been discussed in both literature and within the interviews. The main addition of the interviews is to get an insight into the Dutch tender practices. Taking into account the availability of the data, whether the features are generic and the number of occurrences in both literature and interviews and quantifiability of data, the tender features in table 15 below have been selected to use as input for the ML model.

Table 15 Causes Dropped Tenders, Source: Own Image

Feature name
Complexity
Form of procurement
Estimate of Project size
Duration
Type of object
Type of client
Number of tenderers
Type of contract

Using the model's predictions, it has been possible to determine which features influence the tender price the most. The number of features used as input is optimised, to determine the optimal configuration of input features.

9.1.2. SQ2: What Machine Learning Algorithms are Most Suitable, Taking into Account the Available Data of the Contractor?

Three algorithms are used to develop three separate 'Tender Price Predictor' models. Tender data from a Dutch contractor is used as input for the models. Initially, the tender dataset consisted of 2796 individual tenders. After data-cleaning and outlier detection only 222 tenders were deemed as usable input for the ML model, with three numerical features and six categorical features per tender. An overview of the causes of dropped tenders can be found in figure 12 below.

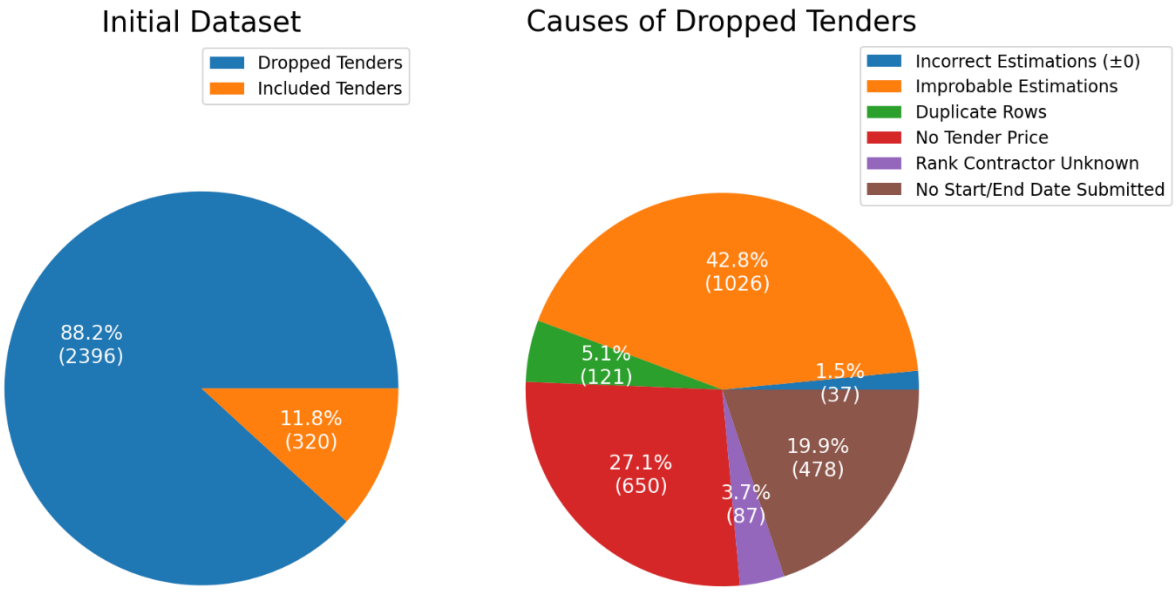


Figure 12 Causes Dropped Tenders, Source: Own Image

The main criteria used to select the algorithms 'interpretability' and 'accuracy'. The models scoring best on these criteria are Linear Regression, DTR and SVR. Linear Regression and DTR are easy to interpret by deriving the feature coefficients, SVR should provide the highest accuracy of the three according to literature.

9.1.3. SQ3: How Accurate Are Tender Price Predictions by Applying Machine Learning Algorithms Using Historical Project Data?

Three ML models have been developed to predict the tender's price. Of the three models, the SVR model performed the best with an R-Squared of 0.846. The SVR model includes an optimised set of features, which is a subset of the initial dataset. The following features influence the tender's price the most:

Table 15 Optimal Subset of Features

Feature Name	Feature Importance
Estimate	0.866
Sqrt_Duration	0.051
Contract_RAW	0.053
Procurement_Price-Only	-0.088

The initial estimate is considered to be significantly more important than the other features. This indicates that the initial estimate influences the tender price more than the duration, type of contract and procurement form. The positive signs for the duration and type of contract state that an increase of the duration or the inclusion of a RAW-contract increase the tender's price compared to projects where this is

not the case. The negative sign for 'Procurement_Price-Only' implies that tenders that are procured only on the lowest price, tend to have a lower tender price compared to EMVI-tenders.

Comparing the model's predictions to the actual tender price, a mean absolute percentage error (MAPE) of 23.5 % is obtained which is equivalent to an accuracy of 76.5%. This value is marginally lower than the MAPE of the experts' estimations, which obtain a MAPE of 23.3 %. Although the absolute mean deviation of the model's predictions is larger than the experts' estimations, the majority of the model's predictions are more accurate than the expert's estimations (53,6% vs 46,4%). Also, the average absolute price deviation in € of the model is lower than the experts' estimations (125.581 € vs 141.964 €).

9.1.4. SQ4: How Can the Tender Price Predictor Be Used Effectively Within the Tender Practices of Dutch Contractors?

Generally speaking, the interviewees agree that contractors may benefit from the Tender Price Predictor when it is possible to accurately predict, within a relative maximum deviation of ±30% (subsection 7.2.1), the winning tender price. The use-case of the Tender Price Predictor may be improved when project-specific or sector-specific characteristics would be used in order to meet the demands of more specialised sectors within the contractor's organization(subsection 7.2.3). Also, the EMVI-component and complexity feature could be improved upon by including more quality-plan components and replacing the tender category feature by more specific complexity variables (subsection 7.2.2).

The Tender Price Predictor could encounter obstacles upon implementation. Requiring too much effort from the users may result in worse quality of data. Both the users of the model and the managers submitting tender data should be trained accordingly in order to obtain maximum efficiency (subsection 7.2.4).

9.1.5. RQ: How Can a Machine Learning Algorithm, Predicting the Tender's Price Using Tender Project Data, Be Developed to Support the Contractor's Decision to Tender?

ML models may aid the contractor during the tender procedure of infrastructure projects at two given moments, according to the interviewed professionals. Besides substantiating the decision-to-tender during the first stage gate, ML models may also aid tender managers during later stages of the tender. For example, proposed implementations are during the design of the plans, fuelling brainstorms or by substantiating decision-making by objective predictions instead of relying on gut feeling or experience of tender professionals.

Three models have been selected as potential Tender Price Predictors: Linear Regression, DTR and SVR. The models have been trained to predict the winning tender price in euros, using 222 tenders with 8 tender features per tender as input: 1) Complexity, 2) Form of Procurement, 3) Estimate of Project Size, 4) Duration, 5) Type of Object, 6) Type of Client, 7) Number of Tenderers and 8) Type of Contract.

Of the three models, the SVR model performed the best with an R-Squared of 0.846. The SVR model includes an optimised set of features, which is a subset of the initial dataset. The SVR model outperforms both the Linear Regression model and the DTR model.

Comparing the model's predictions to the actual tender price, a mean absolute percentage error (MAPE) of 23.5% is obtained which is equivalent to an accuracy of 76.5%. This value is marginally lower than the MAPE of the experts' estimations, which obtain a MAPE of 23.3 %. This appears to be caused by an incorrectly categorised datapoint. Modifying this datapoint improves the model's MAPE from 23.5 % to 22.0 %. Although the absolute mean deviation of the model's predictions is larger than the experts' estimations, the majority of the model's predictions are more accurate than the expert's estimations (53,6% vs 46,4%). Also, the average absolute price deviation in € of the model is lower than the experts' estimations (125.581 € vs 141.964 €).

The model could be improved in the future by modifying the model into a sector-specific Tender Price Predictor, e.g. focusing on solely road construction, rail infra or other civil engineering disciplines. To achieve this, more sector-specific features should be accounted for as input.

In order for Dutch contractors to start considering predictive models like the Tender Price Predictor, plans should be worked out on who will be using to use the tool, where will the data come from and how can a high quality of data be guaranteed. Lots of useful information on tenders has been deemed unusable due to missing values, wrong input and features not being represented in databases.

Predictive models like the Tender Price Predictor may aid contractors in the near future by improving the quality and accuracy of tender price estimating by incorporating objective data on relevant projects of the past. However, in order for such models to be used effectively, a higher quality of data on tenders should be guaranteed in the contractor's information systems.

9.2. Recommendations for Practice

The majority of the initial tender dataset was deemed unusable, with barely 10% making the final dataset. This was caused by a variety of reasons, but the main causes are improbable estimations (42.8 %), no start date or end date submitted (19.9 %) and no tender price submitted (27.1 %). All three causes may be fixed by a higher quality of input contributions of the responsible tender managers. Improbable estimations are initial estimations that are equal to the final tender price, which is highly unlikely given that the estimations are made in an earlier stage of the tender process. It seems that these values are entered after completing the tender. This may be prevented by obliging the responsible tender professionals to fill the cells during the specified tender stages. Also, it should be emphasized that the quality of these entered values do impact the usability of the data to prevent the inputs are rushed.

It was surprising to discover that it was unclear who kept track of what data upon investigating where data on tenders could be found. Eventually, data was derived from the central database mainly used by the contractor's tender board. At the same time, tender strategists kept track of similar data on tenders they had worked on in the past. It would be more efficient for the organization to centralise these databases and be more transparent to reduce these inefficiencies. Also, in order to improve the contractor's understanding of how their data may be used to improve their tender process, it is recommended to hire data analysts or data scientists.

Taking this research as a proof of concept, focusing on specific sectors within the contractor's organization instead of focusing on all infra projects in the database would improve the potential of a successful Tender Price Predictor. The current tender predictor makes use of various general, non-project-specific features. BAM could benefit from investigating how such predictive models perform when modified such that it fits specific sectors or branches of the organization (e.g. road construction or rail construction). Such models are able to include more project-specific elements like material quantities, typical complexity elements inherently connected with the sector or other tender price influencing features.

9.3. Recommendations for Further Research

It would be interesting to investigate whether similar models could be designed for the prediction of the eventual project's result after winning a tender, instead of the tender price. This could either be done by means of a regression model (to predict a numerical outcome) or by a classification algorithm predicting whether a tender may be profitable or not. Winning the tender itself does not produce a positive cash flow unless its actual costs remain lower than the budgeted costs, so focusing on this aspect would be of greater value than predicting the tender's price.

This research focused on predicting the tender's price of Dutch infrastructure projects. Dutch infrastructure projects are characterised by the EMVI-procurement principle. It is recommended to replicate this research study for tender practices in foreign countries. The purpose of this replication is to see whether different tender features would be more dominant in such cases.

The current models are designed to be generic, including tenders from all civil engineering sectors of the Dutch contractor. As a result, the tender feature 'Estimate' was included to have an initial indication of the size of the project. The 'Estimate'-feature, however, appeared to be the most dominant tender feature of all upon the development of the model. To decide which tender features are most important, similar studies should be conducted in specific sectors instead of looking at all sectors at once to include sector-specific elements like material quantities, typical complexity elements inherently connected with the sector or other tender price influencing features.

9.4. Reflection

Working individually on a research topic for over 7 months was a new experience for me. Whereas I previously thought that writing a master thesis required the researcher to have a tunnel-focus on their thesis, I discovered that this is the complete other way around: besides the steps, you complete towards answering your research question, you also have to consider your planning and the planning of others in the organization into account. This research combines a number of different research methods, including a comparative literature study, two interview rounds with experts and the development of three Machine Learning models. All these aspects of research were new to me, and quite the challenge.

This thesis embodies all the different types of new skills that I have acquired in the past seven months. As a 'Construction Management and Engineering'-student with a fascination for data, the opportunity provided by BAM Infra to combine both data engineering and construction management was the perfect combination. From the start on I had the idea that I could completely lose myself in this topic, sometimes a bit too much in hindsight.

Machine Learning and data engineering have been new topics for me. I learned basic programming in Matlab during my bachelor and chose an elective on Python programming last year to prepare myself on a coding-related master thesis. The research requiring me to learn plenty of new skills, combined with the COVID-pandemic, resulted in me losing the overview of my graduation thesis while surviving the lockdowns in my 9.0 m^2 dorm. I am still very grateful to my supervisors for thinking along to improve my focus again, as being able to work again in BAM's office or on the TU Delft campus improved my work ethic once again.

One of the main aspects of my research process that I would have done differently with the knowledge of now is the collection of data and development of the Machine Learning model. The first 1.5 months of my research were devoted to literature study, and literature study only. Upon completion of my literature study and the expert interviews, it suddenly became clear to me that gathering data on the tender features was not as straightforward as I initially thought. This resulted in a delay of the model development, with about 2 months of data-cleaning and coding of the algorithms compared to the initial 4 months. So instead of focusing solely on single tasks, I would and should have conducted more research-related activities in parallel if could do it over again.

References

Agerberg, John-niclas. 2012. "Risk Management in the Tendering Process." *Journal of Economics and Business* 2 (3): 52–59.

Akhilendra. 2021. "Evaluation Metrics for Regression Models | MAE vs MSE vs RMSE vs RMSLE." 2021. https://akhilendra.com/evaluation-metrics-regression-mae-mse-rmse-rmsle/.

Akintoye, A S. 1999. "Factors Influencing the Project Cost Estimating Decision." *CIB W55 & W65 Joint Triennial Symposium Customer Satisfaction : A Focus for Research & Practice*, no. September: 8.

Akintoye, Akintola, and Eamon Fitzgerald. 2000. "A Survey of Current Cost Estimating Practices in the UK." *Construction Management and Economics* 18 (2): 161–72. https://doi.org/10.1080/014461900370799.

Alsaawi, Ali. 2016. "A Critical Review of Qualitative Interviews." *SSRN Electronic Journal* 3 (4): 149–56. https://doi.org/10.2139/ssrn.2819536.

Atkinson, Roger. 1999. "Project Management: Cost, Time and Quality, Two Best Guesses and a Phenomenon, Its Time to Accept Other Success Criteria." *International Journal of Project Management* 17 (6): 337–42. https://doi.org/10.1016/S0263-7863(98)00069-6.

Awad, and Khanna. 2015a. "Support Vector Regression." In *Efficient Learning Machines*, 67–80.

Awad, Mariette, and Rahul Khanna. 2015b. *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. https://doi.org/10.1007/978-1-4302-5990-9.

Bakshi, Chaya. 2020. "Random Forest Regression." 2020. https://levelup.gitconnected.com/random-forest-regression-209c0f354c84.

Başţanlar, Yalin, and Mustafa Özuysal. 2014. *MicroRNA Target and Gene Validation in Viruses and Bacteria*. *Methods in Molecular Biology*. Vol. 1107. https://doi.org/10.1007/978-1-62703-748-8_13.

Bhattacharyya, Saptashwa. 2018a. "Towards Data Science | Ridge and Lasso Regression: L1 and L2 Regularization." 2018. https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b.

———. 2018b. "Towards Data Science | Ridge and Lasso Regression: L1 and L2 Regularization." 2018.

Bonaccorso, Guiseppe. 2017. *Machine Learning Algorithms*. Vol. 1.

Breiman, Leo. 2019. "Random Forests." *Machine Learning* 45 (1): 5–32.

Brook, Martin. 2008. *Estimating and Tendering for Construction Work: Fourth Edition*. *Estimating and Tendering for Construction Work: Fourth Edition*. https://doi.org/10.4324/9780080878102.

Brownlee. 2020a. "Machine Learning Mastery | How to Compare Machine Learning Algorithms in Python with SciKit-Learn." 2020.

———. 2020b. "Machine Learning Mastery | How to Configure K-Fold Cross-Validation." 2020. https://machinelearningmastery.com/how-to-configure-k-fold-cross-validation/.

Brownlee, Jason. 2020c. *Data Preparation for Machine Learning - Data Cleaning, Feature Selection, and Data*. https://doi.org/10.1109/ICIMCIS51567.2020.9354273.

Chao-Duivis, M. A. B. 2019. *Proportionality Guide*. https://www.pianoo.nl/sites/default/files/media/documents/proportinality-guide-Engels-1st-revision-april2016.pdf.

Chen, Qi, Zhaopeng Meng, Xinyi Liu, Qianguo Jin, and Ran Su. 2018. "Decision Variants for the Automatic Determination of Optimal Feature Subset in RF-RFE." *Genes* 9 (6). https://doi.org/10.3390/genes9060301.

Cheng, Min Yuan, and Yu Wei Wu. 2009. "Evolutionary Support Vector Machine Inference System for Construction Management." *Automation in Construction* 18 (5): 597–604. https://doi.org/10.1016/j.autcon.2008.12.002.

Cortes, Corinna, and Vladimir Vapnik. 1995. "Support-Vector Networks." *Machine Learning* 20: 273–97. https://doi.org/10.1111/j.1747-0285.2009.00840.x.

Dhiraj, K. 2019. "Top 5 Advantages and Disadvantages of Decision Tree Algorithm." 2019. https://dhirajkumarblog.medium.com/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a.

———. 2020a. "Random Forest Algorithm Advantages and Disadvantages." 2020. https://dhirajkumarblog.medium.com/random-forest-algorithm-advantages-and-disadvantages-1ed22650c84f.

———. 2020b. "Random Forest Algorithm Advantages and Disadvantages." 2020.

Dreschler, Marco. 2009. *Fair Competition: How to Apply the 'Economically Most Advantageous Tender'(EMAT) Award Mechanism in the Dutch Construction Industry*. http://www.narcis.nl/publication/RecordID/oai:tudelft.nl:uuid:e3c7f772-adc1-477c-9341-43a2d68675df.

Drucker, Burges, Kaufman, Smola, and Vapnik. 1996. "Advances in Neural Information Processing Systems 9." In *Support Vector Regression Machines*. https://books.google.nl/books?hl=nl&lr=&id=QpD7n95ozWUC&oi=fnd&pg=PA155&dq=support+vector+regression+advantages&ots=iEmwpDXSew&sig=_e2Q3V42jKdvIIxYYL9RZjvKLOM&redir_esc=y#v=onepage&q&f=false.

Elhag. 2002. "Tender Price Modelling: ANNs and Regression Techniques Taha."

Elhag, and Boussabaine. 1998. "An Artificial Neural System for Cost Estimation of Construction Projects." *Proceedings of the 14th Annual ARCOM Conference* 1 (September): 219–26. http://www.arcom.ac.uk/-docs/proceedings/ar1998-219-226_Elhag_and_Boussabaine.pdf.

Elhag, Boussabaine, and Ballal. 2005. "Critical Determinants of Construction Tendering Costs: Quantity Surveyors' Standpoint." *International Journal of Project Management* 23 (7): 538–45. https://doi.org/10.1016/j.ijproman.2005.04.002.

Elite Data Science. 2019. "Modern Machine Learning Algorithms: Strengths and Weaknesses." 2019. https://elitedatascience.com/machine-learning-algorithms.

———. 2021. "Elite Data Science | Modern Machine Learning Algorithms: Strengths and Weaknesses." 2021. https://elitedatascience.com/machine-learning-algorithms.

Elumalai, Goku. 2019. "Pros and Cons of Common Machine Learning Algorithms." 2019. https://medium.com/@gokul.elumalai/pros-and-cons-of-common-machine-learning-algorithms-45e05423264f.

Emerson, Sophie, Ruairi Kennedy, Luke O'Shea, and John O'Brien. 2019. "Trends and Applications of Machine Learning in Quantitative Finance." *8th International Conference on Economics and Finance Research (ICEFR).*

European Commission. 2021a. "Applying EU Law." 2021. https://ec.europa.eu/info/law/law-making-process/applying-eu-law_en.

———. 2021b. "Types of EU Law." 2021. https://ec.europa.eu/info/law/law-making-process/types-eu-law_en.

European Union. 2014a. "Directive 2014/23/EU of the European Parliament and of the Council of 26 February 2014 on the Award of Concession Contracts." *Official Journal of the European Union* L 94/1 (28.3.2014): 1–64.

———. 2014b. "Directive 2014/24/EU of The European Parliament and of The Council of 26 February 2014 on Public Procurement and Repealing Directive 2004/18/EC (Text with EEA Relevance)." *Brussels Commentary on EU Public Procurement Law* 2014 (January 2014): 65–242. https://doi.org/10.5040/9781509923205.0008.

———. 2014c. "DIRECTIVE 2014/25/EU OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 26 February 2014 on Procurement by Entities Operating in the Water, Energy, Transport and Postal Services Sectors and Repealing Directive 2004/17/EC (Text with EEA Relevance)." *Brussels Commentary on EU Public Procurement Law*, no. March 2004: 243–374. https://doi.org/10.5040/9781509923205.0014.

———. 2021. "Internal Market, Industry, Entrepreneurship and SMEs." 2021. https://ec.europa.eu/growth/single-market/public-procurement/rules-implementation/thresholds_en.

Flom, Peter. 2018. "The Disadvantages of Linear Regression." 2018. https://sciencing.com/disadvantages-linear-regression-8562780.html.

Flyvbjerg, Skamris, and Buhl. 2004. "What Causes Cost Overrun in Transport Infrastructure Projects?" *Transport Reviews* 24 (1): 3–18. https://doi.org/10.1080/0144164032000080494.

Fradkov, Alexander L. 2020. "Early History of Machine Learning." *IFAC-PapersOnLine* 53 (2): 1385–90. https://doi.org/10.1016/j.ifacol.2020.12.1888.

Fumo, David. 2017. "Types of Machine Learning Algorithms You Should Know." 2017. https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861.

"Geeks for Geeks | Decision Tree Regression Using SKLearn." 2021. 2021. https://www.geeksforgeeks.org/python-decision-tree-regression-using-sklearn/.

Gondia, Ahmed, Ahmad Siam, Wael El-Dakhakhni, and Ayman H. Nassar. 2020. "Machine Learning Algorithms for Construction Projects Delay Risk Prediction." *Journal of Construction Engineering and Management* 146 (1): 04019085. https://doi.org/10.1061/(asce)co.1943-7862.0001736.

Gurucharan. 2020. "Towards Data Science | Machine Learning Basics: Decision Tree Regression." 2020. https://towardsdatascience.com/machine-learning-basics-decision-tree-regression-1d73ea003fda.

Hashemi, Ebadati, and Kaur. 2020. "Cost Estimation and Prediction in Construction Projects: A Systematic Review on Machine Learning Techniques." *SN Applied Sciences* 2 (10): 1–27. https://doi.org/10.1007/s42452-020-03497-1.

Haykin, Simon. 2009. "Rosenblatt ' s Perceptron." *Neural Networks and Learning Machines*, no. 1943: 47–67.

Hussain, Kashif, Mohd Najib, Mohd Salleh, Shabana Talpur, and Noreen Talpur. 2018. "Big Data and Machine Learning in Construction: A Review." *International Journal of Soft Computing and Metaheuristics* x (x): 1–9. file:///C:/Users/iko/Downloads/BigDataandMachineLearninginConstruction-AReview-RG.pdf.

Jain, Murty, and Flynn. 1999. "Data Clustering." *ACM Computer Surveys* 31 (3): 264–323. https://doi.org/10.4018/978-1-5225-1776-4.ch002.

Kamiri, Jackson, and Geoffrey Mariga. 2021. "Research Methods in Machine Learning: A Content Analysis." *International Journal of Computer and Information Technology(2279-0764)* 10 (2). https://doi.org/10.24203/ijcit.v10i2.79.

Kossiakoff, Alexander, William N. Sweet, Samuel J. Seymour, and Steven M. Biemer. 2011. *Systems Engineering Principles and Practice: Second Edition*. *Systems Engineering Principles and Practice: Second Edition*. https://doi.org/10.1002/9781118001028.

Krueger. 2021. "Towards Data Science | LASSO Increases the Interpretability and Accuracy of Linear Models." 2021. https://towardsdatascience.com/lasso-increases-the-interpretability-and-accuracy-of-linear-models-c1b340561c10.

Kultin, N B, D N Kultin, and R V Bauer. 2021. "Application of Machine Learning Technology to Analyze the Probability of Winning a Tender for a Project" 32 (2): 29–33. https://doi.org/10.15514/ISPRAS.

Lewis, Roger J, D Ph, and West Carson Street. 2000. "An Introduction to Classification and Regression Tree (CART) Analysis." *2000 Annual Meeting of the Society for Academic Emergency Medicine*, no. 310: 14p.http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.95.4103&rep=rep1&type=pdf.

Matel, Erik, Faridaddin Vahdatikhaki, Siavash Hosseinyalamdary, Thijs Evers, and Hans Voordijk. 2019. "An ANN Approach for Cost Estimation of Engineering Services." *International Journal of Construction Management* 0 (0): 1–14. https://doi.org/10.1080/15623599.2019.1692400.

Metwalli, Sara. 2020. "Towards Data Science | How to Choose the Right Machine Learning Algorithm for Your Applicationitle." 2020. https://towardsdatascience.com/how-to-choose-the-right-machine-learning-algorithm-for-your-application-1e36c32400b9.

Ministry of Economic Affairs and Climate Policy. 2018. "Procurement Monitoring Report of the Netherlands."

Mitchell, Tom M. 1997. *Machine Learning*. https://doi.org/10.1109/ICDAR.2019.00014.

———. 2006. "The Discipline of Machine Learning." *Machine Learning* 17 (July): 1–7. http://www-cgi.cs.cmu.edu/~tom/pubs/MachineLearningTR.pdf.

Mitsa, Theophano. 2019. "Towards Data Science | How Do You Know You Have Enough Training Data?" 2019. https://towardsdatascience.com/how-do-you-know-you-have-enough-training-data-ad9b1fd679ee.

Morgan, Jake. 2014. "Classification and Regression Tree Analysis." *Technical Report No.1*.

Müller, Andreas, and Sarah Guido. 2017. *Introduction to Machine Learning with Python*. 4th ed. O'Reilly.

Odusami, Koleola T., and Henry N. Onukwube. 2008. "Factors Affecting the Accuracy of Pre-Tender Cost Estimate in Nigeria." *COBRA 2008 - Construction and Building Research Conference of the Royal Institution of Chartered Surveyors*.

Overheid.nl. 2012. "Aanbestedingswet 2012." 2012. https://wetten.overheid.nl/BWBR0032203/2019-04-18.

Pant, Ayush. 2019. "Towards Data Science | Introduction to Linear Regression and Polynomial Regression." 2019. https://towardsdatascience.com/introduction-to-linear-regression-and-polynomial-regression-f8adc96f31cb.

Pedregosa, Fabian, Olivier\ Grisel, Ron Weiss, Alexandre Passos, Matthieu Brucher, Gael Varoquax, Alexandre Gramfort, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30.

PIANOO. 2021a. "Fase 1: Voorbereiden Opdracht." 2021. https://www.pianoo.nl/nl/inkoopproces/fase-1-voorbereiden-opdracht.

———. 2021b. "Fase 2: Doorlopen Aanbestedingsprocedure." 2021. https://www.pianoo.nl/nl/inkoopproces/fase-2-doorlopen-aanbestedingsprocedure.

———. 2021c. "Fase 3: Uitvoeren Inkoopopdracht." 2021. https://www.pianoo.nl/nl/inkoopproces/fase-3-uitvoeren-inkoopopdracht.

———. 2021d. "Legal Framework Procurement in The Netherlands." 2021. https://www.pianoo.nl/en/legal-framework-procurement-netherlands.

PMBOK 6th Edition. 2017. *Project Management Institute*. 6th ed. Vol. 34. Project Management Institute.

Pollack, Julien, Jane Helm, and Daniel Adler. 2018. "What Is the Iron Triangle, and How Has It Changed?" *International Journal of Managing Projects in Business* 11 (2): 527–47. https://doi.org/10.1108/IJMPB-09-2017-0107.

Raj, Aschwin. 2020a. "Towards Data Science | Unlucking the True Power of Support Vector Regression." 2020. https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0.

———. 2020b. "Towards Data Science | Unlucking the True Power of Support Vector Regression." 2020.

Rane. 2018. "Towards Data Science | The Balance: Accuracy vs. Interpretability." 2018. https://towardsdatascience.com/the-balance-accuracy-vs-interpretability-1b3861408062.

Rijksoverheid. 2021. "Aanbestedingsprocedures Voor Overheidsopdrachten." 2021. https://www.rijksoverheid.nl/onderwerpen/aanbesteden/aanbestedingsprocedures.

Rijt, Jeroen Van de, Michael Hompes, and Sicco Santema. 2010. "The Dutch Construction Industry: An Overview and Its Use of Performance Information." *Journal for the Advancement of Performance Information and Value* 2 (1): 33. https://doi.org/10.37265/japiv.v2i1.117.

Roozenburg, N.F.M., and J. Eekels. 1995. *Product Design: Fundamentals and Methods*. Wiley.

Sajee. 2020. "Towards Data Science | Model Complexity, Accuracy and Interpretability." 2020. https://towardsdatascience.com/model-complexity-accuracy-and-interpretability-59888e69ab3d.

"SciKit-Learn 1.0.1 | Cross-Validation." 2021. 2021. https://scikit-learn.org/stable/modules/cross_validation.html.

"SciKit-Learn 1.0.1 | Sklearn.Model_selection.GridSearchCV." 2021. 2021. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html.

"SciKit-Learn 1.0.1 | Sklearn.Svm.SVR." 2021. 2021. https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html.

"SciKit-Learn 1.0.1 | Sklearn.Tree.DecisionTreeRegressor." 2021. 2021. https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html.

Seidu, R D, B E Young, J Clack, and ... 2020. "Innovative Changes in Quantity Surveying Practice through BIM, Big Data, AI and Machine Learning." … *Journal of Natural* … 4 (1): 37–47. https://www.semanticscholar.org/paper/Innovative-changes-in-Quantity-Surveying-Practice-Seidu-Young/9813d313101978473212e5d315c39e7779a55abf.

Sneiderman, Robby. 2020. "Towards Data Science | From Linear Regression to Ridge Regression, the Lasso, and the Elastic Net." 2020. https://towardsdatascience.com/from-linear-regression-to-ridge-regression-the-lasso-and-the-elastic-net-4eaecaf5f7e6.

Stiti, Karim, and Shih Jung Yape. 2019. "Bid Forecasting in Public Procurement," 1–2, 61–73.

Sutton, Richard, and Andrew Barto. 2014. *Reinforcement Learning: An Introduction*. https://doi.org/10.4018/978-1-60960-165-2.ch004.

Taunk, Dhaval. 2020. "L1 vs L2 Regularization: The Intuitive Difference." 2020. https://medium.com/analytics-vidhya/l1-vs-l2-regularization-which-is-better-d01068e6658c.

Vickery, Rebecca. 2020. "Towards Data Science | The Art of Finding the Best Features for Machine Learning." 2020.

Wang, Yu Ren, Chung Ying Yu, and Hsun Hsi Chan. 2012. "Predicting Construction Cost and Schedule Success Using ANNs Ensemble and Support Vector Machines Classification Models." *International Journal of Project Management* 30 (4): 470–78. https://doi.org/10.1016/j.ijproman.2011.09.002.

Winch. 2020a. "Forming the Project Coalition 5.1 (Chapter 5)." In *Managing Construction Projects*.

———. 2020b. "Forming the Project Coalition 5.1 (Chapter 5)," no. 2009.

Wu. 2020. "Towards Data Science | 3 Best Metrics to Evaluate Regression Models." 2020. https://towardsdatascience.com/what-are-the-best-metrics-to-evaluate-your-regression-model-418ca481755b.

Zhang, Yongcheng, Hanbin Luo, and Yi He. 2015. "A System for Tender Price Evaluation of Construction Project Based on Big Data." *Procedia Engineering* 123: 606–14. https://doi.org/10.1016/j.proeng.2015.10.114.

Appendices

Appendix A – European Public Procurement Directives

Several legal frameworks and regulations are in effect to regulate public procurement both nationally and internationally. Three European Directives on the subject of Public Procurement are currently active. EU Directives are legal frameworks that describe an objective set by the European Union but EU Member States are free to interpret them and to transpose the directives into national law (European Commission 2021b). The European Commission is able initiate infringement procedures when Member States fails to transpose the European Directives into national law (European Commission 2021a).

The EU Directives 2014/23/EU, 2014/24/EU and 2014/25/EU on Public Procurement make sure that public procurement within the European Union and its member states adheres to the general principles of free administration by public authorities and the principle of non-discrimination, equal treatment and transparency (European Union, 2014a, 2014b, 2014c). Active since 2014, all three directives follow European Public Procurement in essence but differ in their focus.

EU Directive 2014/23/EU establishes rules on the award of concession contracts between contracting authorities and contracting entities which pursue rights of exploitation(European Union 2014a). EU Directives 2014/24/EU and 2014/25/EU both establish rules on procedures for public procurement and threshold values on works or services but differ in the involved type of sectors. EU Directive 2014/24/EU applies to classical sectors like construction and infrastructure whereas EU Directive 2014/25EU applies to the special sectors of water, energy and transport (European Union 2014b; 2014c). The Directives provide rules to ensure a transparent procurement process and high value contracts but also include an overview of threshold values. The European Directives apply to tenders or contracts when the value of works or services equal or exceed the thresholds as stated in the Directives or adjusted by the European Commission (European Union 2014b). National law applies if the value of the described works or services is lower than the European thresholds, nevertheless respecting the general principles as of the EU. The original threshold values and adjusted values as of January 1st 2020 can be found in table 21. The EU re-evaluates the threshold values every two years.

Table 20 Procurement Threshold Values

Contract type	**Original threshold value excluding VAT (European Union 2014b)**	**2020-2021 threshold value excluding VAT (European Union 2021)**
Public works contract	5 186 000	5 350 000 €
Public subsidised supply and service contract (Central governmental)	134 000 €	139 000 €
Public supply and service contract (sub-central governmental)	207 000 €	214 000 €
Public service contracts (Social services)	750 000 €	750 000 €

The value of the contract has consequences for the public procurement procedure. European procurement procedures have to be followed if the value of the public works or services contract exceeds the threshold value as stated in the European Directives (Rijksoverheid 2021).

Appendix B - Dutch National Public Procurement Law

Besides the European Directives, Dutch Public Procurement is also regulated within Dutch national for works and services both below and above the European threshold value. The three previously mentioned European Directives regarding Public Procurement have been transposed to the Dutch 'Aanbestedingswet 2012' or Dutch Public Procurement Act 2012 (PIANOO 2021d). Dutch National law includes threshold values on governmental contracts of 1.000.000 € excluding VAT for works and 80.000 € per subdivision of the works (Overheid.nl, n.d., Article 2.18). If contracts do not exceed the Dutch national threshold values, it is not mandatory to publicly procure the contracts.

There are a couple of Dutch procurement documents that complement the Dutch Public Procurement Act. The 'Aanbestedingsbesluit 2016' or Dutch Public Procurement Decree 2016 elaborates further upon the articles of the Public Procurement Act 2012 (PIANOO 2021d). The Dutch Proportionality Guide or 'Proportionaliteitsgids' can be consulted by contracting authorities to provide guidance for "a reasonable application of the principle of proportionality (…) every contracting authority must be able to present reasons for its choice to deviate from these rules, for example when selecting more requirements." (Chao-Duivis 2019). The Proportionality Guide functions as a reference book for contracting authorities on the principle of proportionality, as this has not been detailed in the European Directives. The final complementary procurement document is the Works Procurement Regulations 2016 or 'Aanbestedingsreglement Werken 2016' (PIANOO 2021d). The Works Procurement Regulations 2016 describes what types of procedures apply for contracts that are either below or above the European threshold value for contracts.

Appendix C – Consent Form Protocol

Consent Form for 'Interviews Master Thesis Bent Schleipfenbauer'

Please tick the appropriate boxes	**Yes**	**No**
Taking part in the study		
I have read and understood the study information dated [XX-XX-XXXX], or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	○	○
I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.	○	○
I understand that taking part in the study involves a video-recorded interview which is transcribed as text and anonymised in the graduation thesis. Video recordings are destroyed after transcription.	○	○
Use of the information in the study		
I understand that information I provide will be used as input for the graduation thesis of Bent Schleipfenbauer of the TU Delft.	○	○
I understand that personal information collected about me that can identify me, such as name and address, will not be shared beyond the study team.	○	○
Future use and reuse of the information by others		
I give permission for the anonymised transcripts that I provide to be archived in OneDrive so it can be used for future research and learning.	○	○

Signatures

_____ _____ _____

Name of participant [printed] Signature Date

I have accurately read out the information sheet to the potential participant and, to the best of my ability, ensured that the participant understands to what they are freely consenting.

_Bent Schleipfenbauer___ _____ _____

Researcher name [printed] Signature Date

Appendix D –Transcriptions Exploratory Interviews [CONFIDENTIAL]

Appendix E – Tables of Tender Price influencing Features

Appendix E.1 – Tender Features in Price Estimating Literature

A non-ordered overview of the most important tender price influencing features according to literature can be found in table 41. The columns denote the cited authors, while the rows present the names of tender price influencing features in order of appearance. The table includes the features that are represented in literature at least once. The number in the cell represents the order of importance appointed by the authors themselves.

Table 21 Non-ordered overview of all tender features

Feature no.	Feature name.	Odusami&Onukwube 2008	Elhag & Boussabaine 1998	Elhag et al. 2005	Akintoye 1999	Elhag Thesis 2002
1	Expertise of Consultants	1			23	
2	Information Quality	2		4,5	15	
3	Project team's experience	3		2	13	
4	Tender market / no. of tenderers	4	(1-13)	9		
5	Pre-contract design	5		7	11	
6	Complexity (of D&C)	6		8	1	
7	Availability of supplies	7		17	10	
8	Location (rural / urban / inner city / outskirts)	8		43	9	
9	Form of Procurement	9		21	16	x
10	Method of construction	10		36	4	
11	Project size	11	(1-13)	30	2	1,5
12	Duration	12	(1-13)		16	1,5
13	Anticipated frequency of variations	13		1	22	
14	Type of structure	14	(1-13)	44,5	12	(3-11)
15	Client's financial situation	15		43	6	
16	Site constraints/access	16	(1-13)	6	5	2*(3-11)
17	Type of client	17		16	7	
18	Amount of specialist works	18			20	
19	Buildability	19		18	8	
20	Expected project organization	20			19	
21	Number of project team members	21			24	
22	Amount of floors		(1-13)	27		(3-11)
23	Type of project		(1-13)	44,5		
24	Slope		(1-13)	26		(3-11)

25	Start conditions (greenfield / greyfield)		(1-13)			
26	Ground conditions		(1-13)	14		(3-11)
27	Excavation conditions		(1-13)			
28	Stories / Basement levels			12		(3-11)
29	Market condition				3	
30	Project finance method / funding on time			28,5		
31	Partnering arrangements			41,5		
32	Priority on deadline requirements			3		
33	Clients experience of procuring construction			49		
34	Clients requirements on quality			11		
35	Certainty of project brief			10		
36	Relationship client/contractors			28,5		
37	Variation orders and additional works			4,5		
38	Toughness of requirements (inspection and approval of works)			58		
39	Submission of early proposals			25		
40	Confidence in work force			22		
41	Contractor's financial capability			34		
42	Experience similar projects			13		
43	Current work load			40		
44	Communication levels contractor			49		
45	Estimation techniques			59		
46	Planning capability			20		
47	Productivity effects			35		
48	% main work / % subcontracted work			61,5		
49	No. Subcontractors			64		
50	Mark-up %			66		
51	Record of payments to contractor			66,5		
52	Claims record of contractor			23		
53	Present claims (size and quality)			41,5		
54	Accidents on site record			67		
55	Bond/warranty			63		

	arrangements					
56	CDM regulations			65		
57	Type of foundation			37,5		(3-11)
58	Offsite pre-fabrication			52		
59	Cladding external walls			51		
60	Complexity building services			6		
61	Phasing requirements			15		
62	Quality of finishing			47		
63	Type of contract			46		(3-11)
64	Payment type (fixed, cost plus, etc)			32		
65	Tender selection (open, negotiation, two-stage)			21		
66	Spread of risk between parties (client, consultant, contractors)			31		
67	Claims and disputes resolution methods			55		
68	Interviewing of selected prospective contractors			57		
69	Labour prices/availability			19		
70	Plant prices/availability			39		
71	Weather conditions			56		
72	Government regulations			60		
73	Interest rate			54	(3-11)	
75	Stability			33		
76	Lead time for materials				18	
77	Construction site activities sequencing				21	

Appendix E.2 – Tender Features in Machine Learning Models

Feature no.	**Feature name.**	**Zhang et al. 2015**	**Kultin et al. 2021**	**Matel et al. 2019**
78	Type of station	1		
79	Excavation depth	2		
80	Flat shape	3		
81	Hydrological conditions	4		
26	Soil conditions	5		
8	Surrounding environment	6		
82	Budget		1	
23	Type of work		2	

12	Duration		3	3
83	Number of contracts		4	
4	Number of applications		5	
6	Complexity		6	
84	Estimation of business expenses		7	
85	Estimation of project urgency		8	
86	Project intensity			1
21	Number of project team members			2
63	Contract type			4
87	Project phases			5
11	Project scale			6

Appendix E.3 – Tender Features in Dutch Tender Practice

Feature No.	Feature	No. of Occurences (out of five)	Description
88	Need for work	2	Whether project teams of the organisation currently have no work, as this may drive the contractor to lower its price.
89	Relationship with the client	1	Whether the relationship between the contractor and the client considered to be pleasant, based on previous experiences.
15	Financial status client	1	The financial status of the client may indicate whether the client is able to fulfil its financial obligations.
28	No. of expected competitors	1	The number of expected tenderers increases the competitiveness, with more potential 'prijsduikers' in need of work.
63	Type of contract	4	The type of contract, between client and contractor, of the project.
90	Ceiling price	1	The ceiling price is the maximum tender price. This ceiling price is set by the client.
14	Type of object	3	The type of object; road infrastructure, rail infrastructure, real estate, tunnels, etc.
17	Type of client	1	Whether the client is a governmental party or smaller.
2	Quality of information	1	The completeness and quality of the provided information.
5	Precontract design provided	1	Whether a preliminary design is provided, precontract, by the client.
3	Experience of team.	4	Less experienced of the tender team may win the tender, but may forget aspects resulting in higher actual expenses.
6	Multidisciplinary / complexity	4	The complexity and multidisciplinary character of the project. How complex are the to-be-completed works, and are many disciplines involved?
91	Cables underground	2	Cable infrastructure underground of rail infra and tunnels.
12	Planning	3	The expected duration of the project.
9	Form of Procurement	4	The form of procurement, EMVI or price-only.
92	Indirect costs	1	Indirect, non-material, costs: site costs, staff, design, risk, mark-up.

8	Location	2	The environment of the project, rural or urban.
11	Project size	2	An estimate of the project in size.
93	Product-Market-Combination	1	Whether the contractor has constructed a similar project for the same client before.

Appendix F – Tender Feature Selection

Feature no.	Feature name.	Occurences			Requirement No. 1: Occurences (sum>5)	Requirement No. 2: Generic Feature	Requirement No 3: Data in Database	Included in Final Dataset
		No. in Appendix G1	No. in Appendix G2	No. in Appendix G3				
1	Expertise of Consultants	2			No	Yes	No	No
2	Information Quality	3			No	Yes	No	No
3	Project team's experience	3		4	Yes	Yes	No	No
4	Tender market / no. of tenderers	3	1	1	Yes	Yes	Yes	Yes
5	Pre-contract design	3		1	No	Yes	No	No
6	Complexity	3	1	4	Yes	Yes	Yes	Yes
7	Availability of supplies	3			No	Yes	No	No
8	Location	3	1	2	Yes	Yes	No	No
9	Form of Procurement	4		4	Yes	Yes	Yes	Yes
10	Method of construction	3			No	Yes	No	No
11	Project size	5	1	2	Yes	Yes	Yes	Yes
12	Duration	4	2	3	Yes	Yes	Yes	Yes
13	Anticipated frequency of variations	3			No	Yes	No	No
14	Type of structure	5		3	Yes	Yes	Yes	Yes
15	Client's financial situation	3		1	No	Yes	No	No
16	Site constraints/access	5			Yes	Yes	No	No
17	Type of client	3		2	Yes	Yes	Yes	Yes
18	Amount of specialist works	2			No	Yes	No	No
19	Buildability	3			No	Yes	No	No
20	Expected project organization	2			No	Yes	No	No
21	Number of project team members	2	1		No	Yes	No	No
22	Amount of floors	3			No	Yes	No	No
23	Type of project	2	1		No	Yes	No	No
24	Slope	3			No	No	No	No
25	Start conditions	1			No	Yes	No	No

26	Ground conditions	3	1		No	Yes	No	No
27	Excavation conditions	1			No	No	No	No
28	Stories / Basement levels	2			No	No	No	No
29	Market condition				No	No	No	No
30	Project finance method / funding on time	1			No	Yes	No	No
31	Partnering arrangements	1			No	Yes	No	No
32	Priority on deadline requirements	1			No	Yes	No	No
33	Clients experience of procuring construction	1			No	Yes	No	No
34	Clients requirements on quality	1			No	Yes	No	No
35	Certainty of project brief	1			No	Yes	No	No
36	Relationship client/contractors	1			No	Yes	No	No
37	Variation orders and additional works	1			No	Yes	No	No
38	Toughness of requirements (inspection and approval of works)	1			No	Yes	No	No
39	Submission of early proposals	1			No	Yes	No	No
40	Confidence in work force	1			No	Yes	No	No
41	Contractor's financial capability	1			No	Yes	No	No
42	Experience similar projects	1			No	Yes	No	No
43	Current work load	1			No	Yes	No	No
44	Communication levels contractor	1			No	Yes	No	No
45	Estimation techniques	1			No	Yes	No	No
46	Planning capability	1			No	Yes	No	No
47	Productivity effects	1			No	Yes	No	No
48	% main work / % subcontracted work	1			No	Yes	No	No
49	No. Subcontractors	1			No	Yes	No	No
50	Mark-up %	1			No	Yes	No	No
51	Record of payments to contractor	1			No	Yes	No	No
52	Claims record of	1			No	Yes	No	No

	contractor							
53	Present claims (size and quality)	1			No	Yes	No	No
54	Accidents on site record	1			No	Yes	No	No
55	Bond/warranty arrangements	1			No	Yes	No	No
56	CDM regulations	1			No	Yes	No	No
57	Type of foundation	2			No	Yes	No	No
58	Offsite pre-fabrication	1			No	Yes	No	No
59	Cladding external walls	1			No	Yes	No	No
60	Complexity building services	1			No	Yes	No	No
61	Phasing requirements	1			No	Yes	No	No
62	Quality of finishing	1			No	Yes	No	No
63	Type of contract	2	1	4	Yes	Yes	Yes	Yes
64	Payment type (fixed, cost plus, etc)	1			No	Yes	No	No
65	Tender selection (open, negotiation, two-stage)	1			No	Yes	No	No
66	Spread of risk between parties (client, consultant, contractors)	1			No	Yes	No	No
67	Claims and disputes resolution methods	1			No	Yes	No	No
68	Interviewing of selected prospective contractors	1			No	Yes	No	No
69	Labour prices/availability	1			No	Yes	No	No
70	Plant prices/availability	1			No	Yes	No	No
71	Weather conditions	1			No	Yes	No	No
72	Government regulations	1			No	Yes	No	No
73	Interest rate	2			No	Yes	No	No
75	Stability	1			No	Yes	No	No
76	Lead time for materials	1			No	Yes	No	No
77	Construction site activities sequencing	1			No	Yes	No	No
78	Type of station		1		No	No	No	No

79	Excavation depth		1		No	No	No	No
80	Flat shape		1		No	No	No	No
81	Hydrological conditions		1		No	Yes	No	No
82	Budget		1		No	Yes	No	No
83	Number of contracts		1		No	Yes	No	No
84	Estimation of business expenses		1		No	Yes	No	No
85	Estimation of project urgency		1		No	Yes	No	No
86	Project intensity		1		No	Yes	No	No
87	Project phases		1		No	Yes	No	No
88	Need for work			2	No	Yes	No	No
89	Relationship with the client			1	No	Yes	No	No
90	Ceiling price			1	No	Yes	No	No
91	Cables underground			2	No	No	No	No
92	Indirect costs			1	No	Yes	No	No
93	Product-Market-Combinations			1	No	Yes	No	No

Appendix G –Data Cleaning Steps

The data preparation steps as described in section 5.1 are followed in the upcoming subsections.

- Appendix G.1 illustrates the dimensions of the database exports.
- Appendix G.2 explains the context behind the selected variables.
- Appendix G.3 provides an overview of the raw dataset.
- Appendix G.4 explains the completed data cleaning steps and its results.

Appendix G.1 - Raw Data from Tender Database

Two different datasets of the tender database have been acquired to construct a dataset as input for the Tender Price Predictor. The most important features of each dataset need to be combined based on individual tender reference codes, as information on specific tenders and the performances of competitors are found in different databases.

Export 1: Overview Tender Opportunities

The first data export contains an overview of all potential tender opportunities. The dataset provides information on each potential tender, from small to large. Tender dataset 1 describes a total set of 12.681 tenders on with 42 columns / variables. An overview of the most relevant variables is illustrated in table 9.

Table 22 Overview Data Tender Opportunities

Variables Dataset 1 (Dutch)	**Data Type**	**Non-Empty Entries**
Referentie	Numerical	12681
Status	Categorical	12681
Start-Realisation	Date	12265
End-Project	Date	12296
Contract Scope	Categorical	12626
Contract Type	Categorical	12608
Main Deliverable	Categorical	2597
Business Type	Categorical	11870
Tender Category	Categorical	1906
Price / quality	Categorical	12627
Total Entries		12681

Export 2: Overview Performances Tender Participants

The second data export contains a more selective overview on the performance of each separate competitor per tender. Information is collected on the submitted tender prices per competitor, the submitted tender price of BAM and, in case of EMVI tenders, the fictive discounts provided per submitted tender. Tender dataset 1 describes a total set of 11.794 tenders on with 36 columns / variables. An overview of the most relevant variables is illustrated in table 24.

Table 23 Overview Performances Tender Participants

Variables Dataset 2 (Dutch)	Data Type	Non-Empty Entries
Referentie	Numerical	11794
Estimate	Numerical	11789
Tenderer	Categorical	11794
Price	Numerical	10403
Fictive Discount	Numerical	2729
Total Entries		11794

Appendix G.2 - Raw Tender Data Explanation

This section is devoted to the explanation of the tender variables, and how each variable relates to the tender features. The use of some variables as input data may be more straightforward than others, but which will be explained in subsections 5.2.1 – 5.2.15. The data-entries discussed in this section are raw datapoints, which means that the data is yet to be cleaned. As a result, the final dataset will be much smaller in size. This is a result of the tender features representing just a single column of the final dataset, while the final dataset consists of rows (representing the individual tenders) with no empty columns.

Referentie (Data Type: Numerical)

The 'Referentie' is the reference number of the tenders. This number is used throughout the information systems of BAM, and makes it able to combine the data throughout the datasets. With the use of the Pandas library it is possible to combine the datasets, and create a new dataset which solely includes entries that have values for these variables.

Status Reden (Data Type: Categorical)

Status refers to the current state of the tender process. This variable is used to exclude tenders which are either not finished yet, or have been missed by BAM. Tenders in which BAM did not participate are not useful to incorporate as BAM would not have a tender price for these selected tenders. Tenders which are either won or lost are included in the final dataset.

Start-Realisatie / Einde-Project (Data Type: Date)

Start-Realisation and End-Project, the expected dates of realisation and finalization of the project, are used to determine the expected duration of the project. The duration of the projects, not to be confused with the tender period, is estimated during the initial tender phase. The period is determined by the difference between the initial starting date and the completion date.

Contract Scope / Type (Data Type: Categorical)

Various types of contractual agreements have been made between BAM and the client for the works and services provided. Each type of contract has its own corresponding legal structure and implications, e.g. type of payments, responsibility and agreements. Therefore, the feature 'Type of Contract' has been split in two measurable variables; the scope of the contract, and the type of contract. The exact implications of the separate contracts are beyond the scope of this research study.

Main Deliverable (Data Type: Categorical)

'Main deliverable' denotes the type of object that is delivered for the tender, and is documented explicitly in the tender database. The feature has been added recently to the database, which results in a relatively large amount of missing values.

Business Type (Data Type: Categorical)

The 'Business Type' variable describes the origin of the business or client. What stands out, is that it appears that large public clients e.g. Rijkswaterstaat and Prorail appear to be missing from the data entries. This is most likely a mistake within the system itself, as he input is determined within a drop-down menu. This means that either the cells are missing, or Prorail and Rijkswaterstaat are logged as Governmental parties.

Tender Category (Data Type: Categorical)

The 'Tender Category' is qualitative label given to tenders by BAM. Tenders may be given 5 different types of labels, from 'Category E' up until 'Category A'. Category A is given to the largest, most complex tenders while Category E is given to small, non-complex tenders. The following criteria are taken into account:

- Order Value
- Contract Type
- Contract Experience
- Risks
- Region
- Logistics
- Organization Complexity
- Technological Complexity
- Ground conditions
- Client Track Record
- Client Relationship

The variable Tender Category is taken into account as a degree of total project complexity, as the separate scores on the criteria are not provided in BAM's database. Ideally, technological / organizational are take into account as stand-alone features but this is not possible.

Price / Quality (Data Type: Categorical)

Price / quality denotes the form of procurement of the tender. Tenders are divided in two forms of procurement, either price only or quality i.e. EMVI tenders. In the case of EMVI tenders, fictive discounts are subtracted from the submitted tender price when certain quality criteria are met. This is an important metric to take into account, as the lowest price does not necessarily win the tender until the fictive discount is subtracted from the tender price.

Estimate (Data Type: Numerical)

An estimate of the size of the tender is provided during the early tender phase. This is not a final or conclusive estimate, but the ballpark estimate serves as a general indicator of how large the project is going to be. This tender features provides a starting point of the tender pricing.

Tenderer (Data Type: Categorical)

The tenderer is the name of the corresponding tenderer that submitted the bid. By determining the amount of unique participants that participate in a tender, it is possible to construct a numerical value which denotes the amount of tenderers per tender. In essence the data type is categorical, but this is changed to numerical by counting the amount of unique participants.

Price (Data Type: Numerical)

Price is the submitted tender price of each tenderer. The winning tender price depends on the bid of the competing tenderers, BAM, and the form of procurement. When tenders are procured through the EMVI-procedure, the discounted tender price is taken into account as this is the tender bid that closes the deal. In case of price-only tenders, no quality plans are taken into account resulting in no fictive discounts.

Fictive Discount (Data Type: Numerical)

Fictive discount awarded to tenderers in case of EMVI tenderers. By subtracting the discount of the submitted tender price it is possible to determine the definitive tender price. The tenderer with the lowest 'fictive' tender price is awarded the tender.

Appendix G.3 - Overview Raw Dataset

After selecting the most relevant columns of every dataset, a final dataset is constructed by combining the three dataset exports. The data exports are concatenated on the variable areferentie'. In this case, this means that information of data exports 1 and 3 are 'glued' on data export 2. This results in a dataset containing tender price information of every individual tenderer per tender, combined with more general information of the features in data exports 1 and 3. The final 'raw dataset consists of 11.844 separate tenderers with 25 descriptive columns, of which some are temporary.

As not all datapoints will have values for some columns, so-called 'Not A Number'-values or NaNs for short are used to replace these empty cells. An final overview of the missing values is provided in table 25.

Table 24 Missing Values Dataset

Variable Name Raw Dataset	**Number of Missing Values**
Referentie	0
Estimate	5
Number of Tenderers	0
Tenderer	0
Price	1396
Fictive Discount	9085
Empirics	9094
Status	6692
Name Project	6692
Business_Unit	6693
Start Realisation	6946
End Project	6819
Contract_Scope	6696
Contract_Type	6692
Main_Deliverable	9499
Price / Quality	6692
OG_Type	7176
Tender Category	10400

How to cope with these missing values is discussed in subsection 5.3.3 and 5.3.4, as each variable requires its own type of strategy.

Appendix G.4 - Cleaning the Raw Tender Data

The raw dataset as identified in the previous section is not ready to be used as input for the Tender Predictor. The raw dataset may contain various double entries, double columns, missing values and outliers which may strongly influence the quality of the model's predictions. In order to cope with this, various data cleaning steps are undertaken to improve the quality of the dataset.

Removing duplicate columns

One pair of duplicate columns was found in the raw dataset:

Table 25 Duplicate Columns

Duplicate Column Name	Amount of Missing Values
Estimate_1	6375
Estimate_2	5

The 'Estimate _x' column denotes the estimated size of the tender in €. The large amount of missing values of 'Estimate_1 is a consequence of the concatenation of data exports 1 on data export 2. Estimate_1is dropped from the initial dataset.

Removing duplicate datapoints or rows

For the removal of duplicate datapoints, a distinction has been made between two types of 'duplicate datapoints'. First of all, all duplicate rows are dropped from the dataset. Duplicate rows implies that the entries for all columns of certain datapoints are equal, resulting in identical duplicates.

Second of all, during closer investigation it appeared that some of the estimates in Estimate_2 were identical to the winning tender price. The probability of the early estimate being identical to the winning tender price is extremely low, and results in the Machine Learning model being more positively biased at first sight. This may be a result of the estimated size being filled in after completion of the tender. As this does strongly influence the model's apparent accuracy, all double entries in these columns are dropped. This action significantly influences the size of the dataset, resulting in a decrease of 1189 tenders in which BAM participated.

Besides double entries within the 'Estimate' column, a large number of project entries had unlikely low values. In total, 36 entries included a project estimation of either 0, 1 or 30 €. These values have been excluded from the dataset as well as these are not representative for the actual size of the project.

Marking and Replacing Missing Values

Two types of data are considered during the marking of missing values: numerical and categorical features. Each datatype has its own coping strategy, with numerous coping strategies being available for missing numerical entries. Missing categorical values may be replaced by 'Unknown', as all entries in the columns are strings i.e. in text form. Deleting the entire datapoint due to a missing description would be a waste of data

This differs when the missing values are of the numerical datatype, and depend on the context of the variable. An explanation of coping strategies for numerical features may be found below in table 27.

Table 26 Coping Strategy Missing Values

Feature Name Dataset	Number of Missing Values	Coping strategy	Explanation
Estimate	5	Dropping rows	Rows without an estimate are dropped from the dataset. Replacing by another value, e.g. the mean, may influence the accuracy of the model.
Number of Tenderers	0	-	-
Tender Price	1396	Dropping rows	Rows without the tender price of the project are dropped from the dataset. Without a tender price to predict, it is impossible to include the datapoint in the dataset.
Fictive Discount	9085	Filling with '0'	The absence of a fictional discount in the dataset is not necessarily a problem. Price-only tenders do not have a fictional discount.
Duration	6692	Dropping rows	Rows without the duration of the project are dropped from the dataset. Replacing by another value, e.g. the mean, may influence the accuracy of the model. Rows are dropped without a start or final date, or when the duration is negative.

The 'Duration' feature is engineered by taking the difference between 'Start realisation' and 'End project'. If either column contained a missing value it was still possible to come up with a duration, which was either very large or negative. Rows with zero-entries in either column have been dropped.

After replacing the identified missing values, it is possible to complete the final step of data cleaning.

Identifying outliers by means of statistics

Outliers may influence the model's accuracy, especially in cases of linear regression. To cope with this, statistics are used to identify the outliers. A statistical description of the cleaned numerical variables, including outliers, can be found in table 28. One proxy-variable is added to check whether there are anomalous 'Estimate'-submissions compared to the winning tender price.

Table 27 Stochastic Descriptions Numerical Features

	No. of Tenderers	Tender Price	Duration	Estimate	$Diff = \frac{Estimate - TenderPrice}{TenderPrice}$
Unit	[n]	[€]	Months	[€]	[-]
count	320	320	320	320	320
mean	4.66	2433801.27	8.01	3500213.99	2.32
std	3.12	15235174.63	10.38	26618710.64	-0.97
$CV = \frac{std}{mean}$	0,67	5,89	1,30	7,60	5,53
min	2.00	13900.00	0.03	8578.00	-0.97
25%	3.00	309300.00	2.03	310750.00	-0.02
50%	4.00	626500.00	4.12	630000.00	0.17
75%	5.00	1161749.75	9.84	1046250.00	0.41
max	28.00	259209545.00	85.20	447000000.00	39.82

It is noticed that the coefficient of variation (CV) is relatively high for a number of numerical features. A high CV, the ratio between the standard deviation and the mean, implies that the distribution of the variable is rather skewed with long tails and therefore outliers. The proxy-variable appears to be rather skewed, which means that the ratio between estimate and tender price is extremely high which can also be derived from the max value of 39.82. A max value of 39.82 means that a specific project that the estimate in euros is 40 times as large as the final tender price. Such anomalies may be mistakes regarding input, or

extreme projects. Investigation of the cause remains outside the scope of this research but should be taken into account upon completion.

To cope with these outliers, the Interquartile Range (IQR) method is applied. The IQR-method can applied to get rid of these outliers and decrease the skewness of the distributions. The distance between the first and third quartile of the data points in the datasets, also known as the interquartile range, is the namesake of the method and illustrated in figure 14.

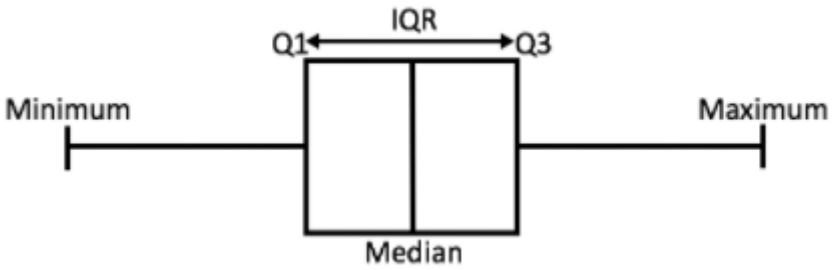


Figure 27 Interquartile Range illustrated, Source: (Chaudhary 2021)

The distance between the Q1 and Q3 values is used to scale the lower and upper bounds of the final dataset to eliminate outliers (Chaudhary 2021). Usually, 1,5 * IQR is either subtracted from Q1 to determine the lower limit and 1,5*IQR is added to Q3 respectively to denote the upper limit of the dataset. After applying the '1,5-IQR method', less skewed distributions for the separate features are obtained.

Data Transformation

The next step in the data preprocessing approach is the transformation of the cleaned variables. Based on the obtained probability density functions of the numerical, the standardization approach as explained in 3.4.2 is chosen. Data transformations in general make for better model performances, when scaled to a similar interval.

The standardization of the features resulted in new distribution functions and boxplots for the corresponding features. Given that the probability distribution functions appear to fit a normal distribution, it may be assumed that the use of the StandardScaler() function is justified. The values are scaled after all preprocessing steps.

Correlation Tender Features and Feature Engineering

Pairwise plots have been generated in order to determine whether features should be modified. These bivariate distributions visualize how the separate features relate to each other. The diagonal entries contain univariate distributions of the variable, as every feature is plotted on both axes. To provide more information to the plot, data-points are coloured based on whether BAM won the tender or not.

Both SVRs and DTRs tend to perform well on non-linear problems. This is not necessarily the case for Linear Regression. The pairwise plot is investigated whether non-linear relations can be identified between the selected output variables and numerical values. This will not be considered for datapoints of one-hot-encoded categorical features as these features have 4 possible locations (any combination of x and y being either 0 or 1). The pairwise plot of the numerical features 'Number of Tenderers', 'Duration' and 'Estimate' together with the output variables of 'Ratio' and 'Tender Price' are provided in figure 15.

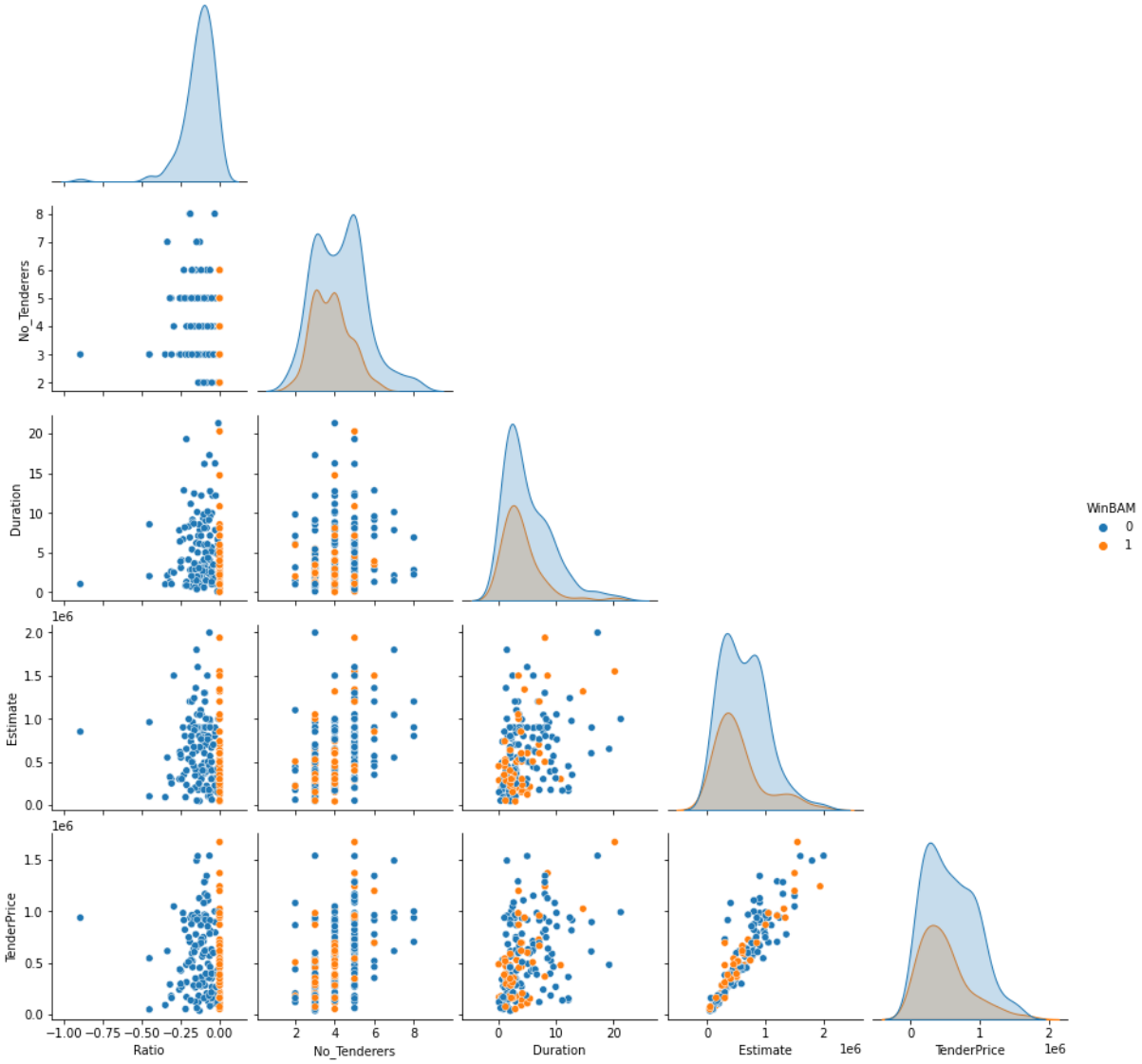


Figure 28 Pairwise Plots Numerical Features, Source: Own Image

It also appears that both the 'Duration' as the 'Number of Tenderers' bivariate distributions with the tender price appear to relate non-linearly. In order to test this, two new features are created by taking the square root of both 'Duration' as the 'Number of Tenders', and it is investigated how these corelate with the tender price.

To investigate the impact of the engineered features, a correlation plot is generated to investigate how the newly constructed features interact with the tender price. Figure 16 on the next page illustrates how the separate features correlate with each other.

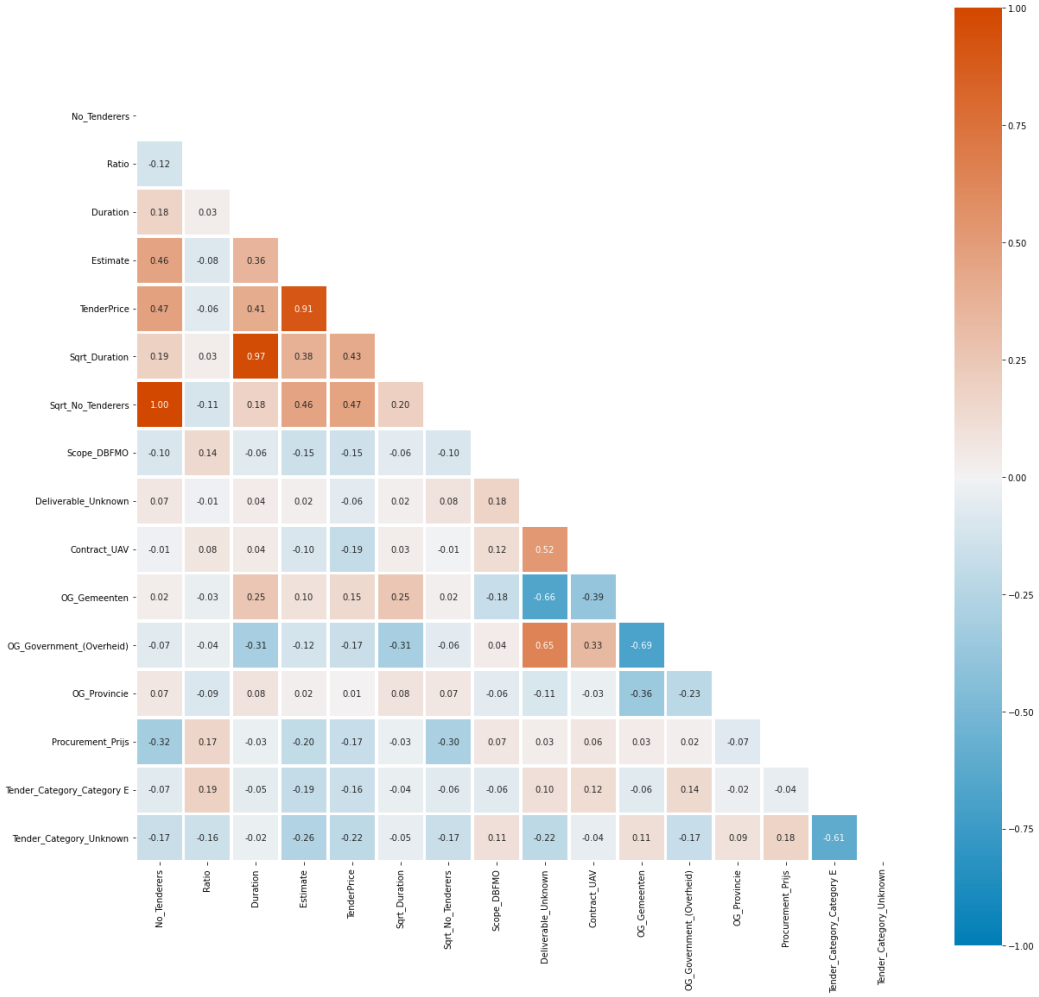


Figure 29 Correlation Plot Dataset, Source: Own Image

Upon investigation, it is noticed that Sqrt_Duration has a stronger correlation with the tender price (0,43) than the original Duration (0,41). The Sqrt_NoTenderers has a slightly weaker correlation (0,44) with the tender price than the original feature (0,45), so the Number of Tenderers features is not modified. Strong correlations exist between the original and modified 'Duration' and 'No_Tenderers' respectively. The weaker correlating features with the TenderPrice are dropped from the dataset.

Also, it should be noted that strong correlation exists between the 'Estimate' and 'TenderPrice' features. This could have been expected, but the correlation is stronger than initially assumed. Recursive Feature Elimination should point out whether the remaining features could improve the accuracy of the model compared to solely using the 'Estimate' feature.

For the remainder, features that appear to have a weak relationship with the tender price are not omitted from the dataset. It may be the case that features one on one do not strongly correlate with the tender price, but they may strongly correlate when combined with other features. Feature set combinations are optimised in subsection 5.6.

Dimensionality Reduction

As discussed in subsection 3.4.2, two popular methods of dimensionality reduction are RFE and PCA. Both types have been considered to reduce the amount of features used as input.

In PCA, PCs are created by combining uncorrelated features and dropping these features from the columns. While it does improve the performance and reduces the computational costs, it does affect the interpretability of the problem. As the original features are combined into new ones, it becomes harder to understand how the original features affect the output variables. This increase in complexity, together with a lack of applicability for categorical features, results in PCA not being considered as an effective dimensionality reduction tool for this specific research.

Recursive feature elimination is implemented in the optimization of hyperparameters. Within the hyperparameter optimization, the amount of selected features is taken into account. The results of this hyperparameter analysis are provided in chapter 6.

Appendix H –Transcriptions Validation Interviews [CONFIDENTIAL]

Appendix I – Decision Tree Visualised (DTR)