Using Machine Learning to Improve the Cost Estimation Process at a Versatile Manufacturing Company

A. van Vlastuin

Faculty of Behavioural, Management and Social sciences (BMS), University of Twente

BSc. Industrial Engineering and Management

Dr. Abhishta & dr. R.A.M.G. Joosten

June 30th, 2023

Acknowledgements

I would like to express my gratitude to my supervisor, dr. Abhishta, for his guidance from start to finish. Thanks to dr. Joosten, also, for ensuring that my academic writing was up to par. Their feedback helped shape this thesis into what it is today.

Thanks goes out to all my colleagues at EeStairs, as well. To Robin, first and foremost, for the data, insights, recommendations and supervision he provided. To all those who participated in the interviews that helped to clarify EeStairs' processes and identify cost drivers, or supported me generally: Cornelis, Harmen, Martijn, Dick, Johan, Rebecca, Sam, Dianne and Mascha.

I am enormously grateful to Daphne for the assistance that allowed me to round off my analysis. Her instructions and critical questions helped me choose what to focus on and what to disregard, putting pen to paper to finish this work.

Lastly, I would like to thank all my friends and family members for their support throughout this process; to Tineke, Titia and Dirk for their close involvement and for offering me a space to write, and finally to Roos whose engagement, advice and sympathy helped me throughout the entire process.

Executive summary

This thesis examines the feasibility of machine learning models to estimate the engineering costs for EeStairs. As is the case in other versatile manufacturing companies (VMCs), their cost estimation process is complicated. Among other factors this is due to the variability of their products, a dominance of manual processes, challenges in linking predicted and actual costs, and a lack of clear specifications provided by the client. To deal with the speed, accuracy and consistency requirements of VMCs' quotation processes, the company identified a need for a time-efficient, objective and interpretable cost estimation method.

Drawing on interviews and existing literature on cost estimation and design & engineering costs, this study identified several cost drivers. Then, data on 53 projects was gathered from EeStairs' databases. Due to output data only being available as an aggregate over projects (while input data was separated by individual staircase and balustrade), heavy feature engineering was conducted. To construct a model that is understandable and implementable by the company, this study focused on regression models and optimised four variants.

Our analysis shows that the Lasso Regression model is best equipped for the task. It performs only slightly better than EeStairs' manual method based on the mean squared error, whereas the percent error is 57% higher. In addition, the model uses only one feature, indicating that many of the engineered features offered little predictive value. This thesis concludes that our model is not accurate enough to offer a reliable improvement over the old method.

Building on the information obtained through this study, we offer a number of recommendations to enable a future machine learning project:

1. Record actual costs and hours in more detail, per balustrade or staircase rather than for the entire project.
2. Use project & product features that are machine learning-usable in the new cost estimation system, to aid data collection and minimise the information loss associated with feature engineering.
3. Implement methods that increase the objectivity and consistency of input and output data, e.g. by firmly distinguishing between the estimated cost and the quoted price or by removing incentives for employees to misreport their hours worked.

All in all, this thesis took the first steps towards a data-driven approach for estimating costs based on project features at the company. We hope this thesis provides a helpful foundation to further automate and improve EeStairs' cost estimation process, as well as those of other versatile manufacturing companies.

Table of Contents

Abbreviations

Abbreviation	Definition
BIM	Building Information Model(ing)
CRM	Customer Relationship Management
DSRM/P	Design Science Research Process/Method
(M)AE	(Mean) Absolute Error
(M)APE	(Mean) Absolute Percentage Error
MSE	Mean Squared Error
SLR	Systematic Literature Review
VMC	Versatile Manufacturing Company

Glossary

Term	Definition
Calculator	Employee who calculates a project’s estimated cost
Quote / quotation	Offer to a customer that describes a project’s details and their associated costs

1 Introduction

In this chapter, we motivate and provide context for the work done in this thesis. Section 1.1 offers a brief overview of the company where the research was executed: EeStairs. After providing the required background, Section 1.2 explains EeStairs' core problem and motivates our research aim. Subsequently, Section 1.3 introduces the scope and requirements of this thesis, and Section 1.4 our methodological framework.

1.1 Problem context

The internship leading to this thesis is conducted at EeStairs, a luxury staircase designer and manufacturer best described as a versatile manufacturing company (VMC), defined by Amaro et al. (1999) as “manufacturers of customised products that are involved in a competitive bidding situation for (nearly) every order received”. Versatility, here, pertains to continually designing and configuring how to manufacture new or modified products, dealing with varying production loads, and dealing with each customer order individually (Kingsman & Souza, 1997).

The company designs, produces and installs made-to-order staircases and balustrades for architects, retailers, offices and private customers. Except for a few standardised products, which are outside the scope of this research, all of EeStairs' projects are one-of-a-kind. Every product is a unique combination of the materials, shapes, finishings and surroundings, some examples of which can be seen in Figure 1.

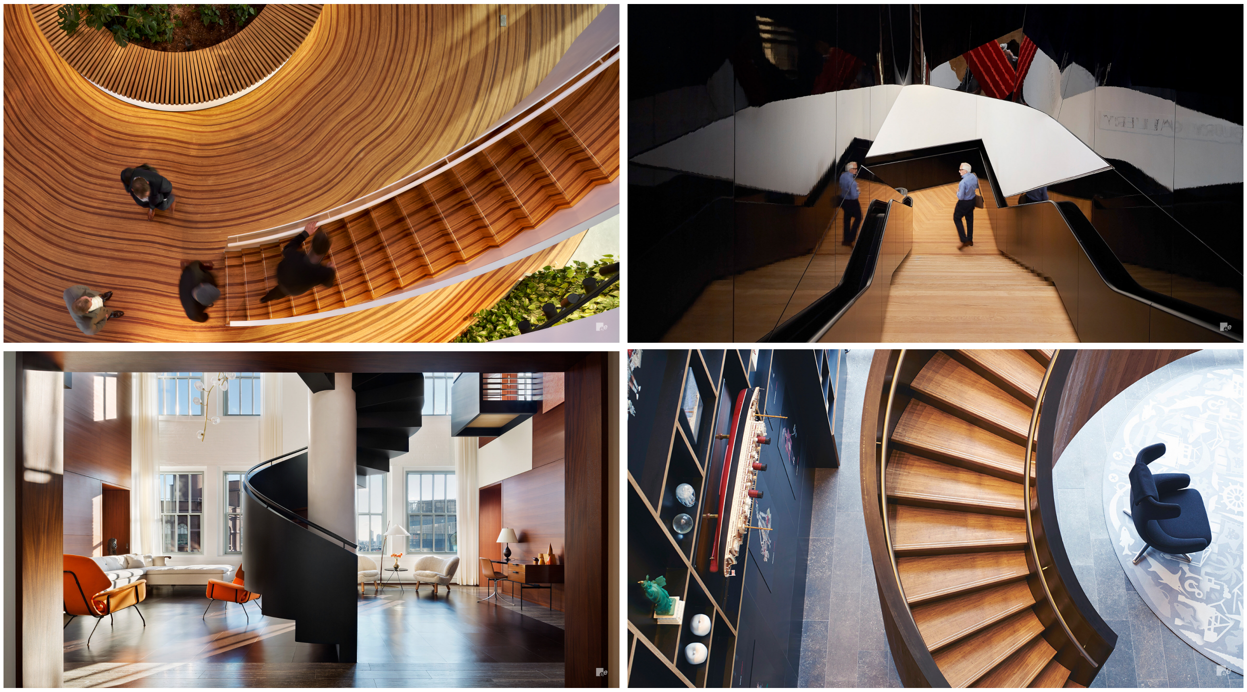


Figure 1: Example staircases and balustrades from EeStairs (photography: Hans Morren).

Figure 1, continued.

1.2 Research objective

EeStairs uses a predominantly manual quotation process and has identified this as a potential bottleneck for future growth. We represent the current quotation process, from estimating a project’s cost to quoting a price to the customer, in Section 2.1. There are three reasons for it being a bottleneck. First, the current quotation process is labour intensive; employees spend several hours at the beginning of every new project making a prediction for the costs, including employee hours. Second, the current process heavily relies on the knowledge of a few experienced employees. This is not scalable, as it is likely these employees will at some point leave the company, or that not all experience can be passed on to new employees. Lastly, the current system is often inaccurate: it regularly over- or underestimates the projects’ cost by dozens of percentage points. For instance, the average percentage error for the engineering cost is 35%.

From Figure 2, we can see most problems stem from two underlying sources: that the quotation process is a manual process, and that EeStairs’ projects are highly customised. The latter is, however, a problem inherent to their business of providing custom-made luxury staircases. Therefore, we will tackle the core problem: the current quotation process is *manual, subjective and experience-based.*

EeStairs’ case is not an isolated one: Kingsman et al. (1996) found that estimating the cost of producing the order and then finding the price to be quoted is a significant problem encountered by VMCs. This is a problem because a poor quality of estimates is often the cause of projects going over the quoted price (Bashir & Thomson, 2001). Additionally, the faster and more accurate the price estimation, the greater the possibility a client accepts the quotation (García-Crespo et al., 2009).

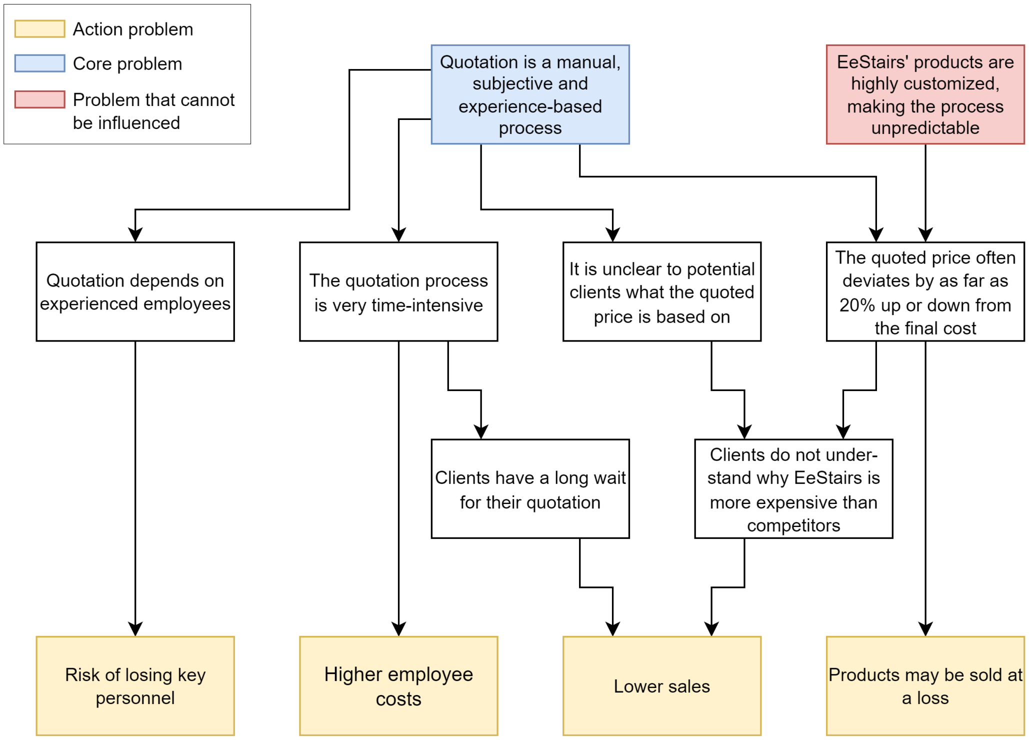

Action problem
Core problem
Problem that cannot be influenced
Quotation is a manual, subjective and experience-based process
EeStairs' products are highly customized, making the process unpredictable
Quotation depends on experienced employees
The quotation process is very time-intensive
It is unclear to potential clients what the quoted price is based on
The quoted price often deviates by as far as 20% up or down from the final cost
Clients have a long wait for their quotation
Clients do not under-stand why EeStairs is more expensive than competitors
Risk of losing key personnel
Higher employee costs
Lower sales
Products may be sold at a loss

Figure 2: Problem cluster using the methodology of Heerkens and Van Winden (2021). We connect EeStairs' concrete (action) problems with the underlying cause, the core problem, which this thesis aims to resolve.

However, the reason VMCs struggle with cost estimation, specifically, is because they often face a unique set of challenges. According to Kingsman et al. (1996), VMCs...

- ...receive one-of-a-kind orders, requiring extensive quotation and some form of activity-based costing.
- ..rely on expert knowledge to estimate costs, basing these estimations on unwritten rules and previous experiences (García-Crespo et al., 2009).
- ...set a price instead of a profit margin, often using rule-of-thumb methods to make adjustments to the estimated cost based on the company, client and market conditions.
- ...have automated some of their processes, while (fixed) capital expenses are harder to allocate to specific projects than (variable) manual labour hours.
- ...heavily rely on bidding requiring speed, accuracy and consistency, factors that (with manual estimation) are to the detriment of each other.
- ...are often unaware of the variance between estimated and actual costs because estimates are discriminated by activity and by product component while the actual costs are recorded by activity and by worker, making it impossible to identify cost variances.
- ...frequently face unclear specifications provided by the client.

Several other authors (e.g. Hvam et al., 2004; Denkena et al., 2009; Zhang et al., 2012) have conducted research into cost estimation for manufacturers of custom-made products. However, these concern companies that either produce at a large scale, produce products with limited variability, or both. Therefore, and because of the unique challenges faced by VMCs, this literature is not directly applicable to EeStairs and companies alike.

This research aims to fill this research gap by estimating the engineering cost for a designer and manufacturer of luxury staircases, hoping to function as a case study for similar companies that wish to improve their quotation process. Moving towards a data-driven approach for estimating the engineering cost for new projects, this study looks at a machine learning approach for predicting engineering costs based on the available historic data. As such, we arrive at the following research question:

Research question:
How can EeStairs use machine learning to improve its cost estimation process?

1.3 Scope & requirements

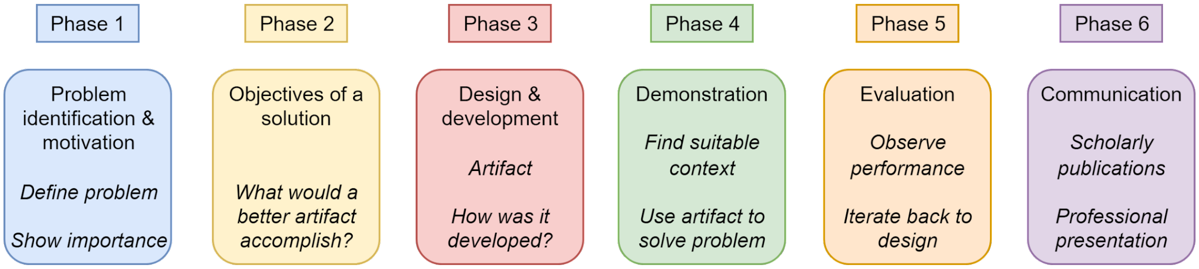
Focus on engineering cost
The company has stated clearly that the quoted price must not just be an aggregate price, but should be made up of the costs of separate parts and activities so customers and employees can understand where the final price comes from. Therefore, and to limit the scope of this thesis, we decided to focus our effort on estimating the engineering cost. Since it is the first step in the production process, it is independent from other activities within a project and thus a good research focus. Engineering is quite complex, embodying nearly all aspects of a product, and thus research into engineering is relatively representative for other parts of the production process: if machine learning works to predict the engineering cost, it should be able to predict other costs as well. This is echoed by Salam and Bhuiyan (2016), who concluded that understanding and being able to estimate the design effort (in terms of person-hours) is crucial in order to estimate a project's cost.

Implementability
EeStairs has expressed the importance of understanding, working with, and, if necessary, improving the cost estimation tool resulting from this research.. As such, our product should be interpretable (as a linear regression formula, for instance). For this reason, we exclude more complex models, such as Neural Networks and Decision Trees, from our research.

1.4 Methodological framework

We apply the design science research process (DSRP) (Peffers et al., 2006) as a methodological framework for this research. According to the authors, DSRP is "a process for carrying out design science research in information systems". The methodology was created to develop technology-based solutions to (business) problems by designing a successful 'design artefact', such as a model or method. Since this research aims to design a tool that improves EeStairs' engineering cost estimation, DSRP seems applicable.

The process comprises six phases illustrated in Figure 3, as based on Peffers et al. (2006).



Phase 1

Problem identification & motivation

Define problem

Show importance

Phase 2

Objectives of a solution

What would a better artifact accomplish?

Phase 3

Design & development

Artifact

How was it developed?

Phase 4

Demonstration

Find suitable context

Use artifact to solve problem

Phase 5

Evaluation

Observe performance

Iterate back to design

Phase 6

Communication

Scholarly publications

Professional presentation

Figure 3: Design Science Research Methodology process model.

The rest of this thesis is structured as follows. Phase 1 of the DSRP is executed in Section 1.1 and Chapter 2, first introducing and then explaining in detail the problem context and available literature. The objectives of a solution are defined in Sections 1.2 and 1.3. Phases 3-6 of the DSRP, then, correspond to the rest of this thesis' chapters. In Chapter 3, we outline the method used to develop our solution, including data collection and model development. In Chapter 4, we present the results, followed by discussion and recommendations in Chapter 5. Chapter 6 offers a conclusion.

2 Literature review and current process description

In this chapter, we provide additional information that will help to understand both the problem and possible solutions. In Section 2.1, we outline EeStairs' current quotation process, including the sales context in which it occurs. In Section 2.2, a systematic literature review is executed to determine the cost drivers of design and engineering. Finally, Section 2.3 outlines the challenges and methods of estimating design costs.

2.1 The quotation process at EeStairs

We establish an understanding of EeStairs' current quotation process based on interviews with directors, engineers and sales employees at EeStairs, together with a document review. What follows is a brief description of this process.

Before a client requests a quotation for the production of their staircase(s) and/or balustrade(s), a design of their requested product has already been made - either by an independent architect, or by EeStairs in what it defines as a (separate & previous) 'design project'. Using this (often quite abstract) design, EeStairs employees the company refers to as 'calculators' then formulate a list of materials, hours, and other associated costs. Over the course of a few hours (or more, depending on the project), they generate a quotation which includes these cost factors and their total sum.

Calculators use a program called TrapCalc to do their calculations with. A longer description of TrapCalc and the associated calculation process can be found in Appendix A, but in brief this software has the following functionalities:

1. It is a framework within which to enter product specifications and estimations.
2. It performs basic arithmetics on estimation.
3. It adds all estimations into a standardised quote in pdf-format.
4. It offers the ability to save estimations made for parts of the staircase for later use.

All other estimations and judgements are performed by calculators themselves. For instance, they add the specifications of a product (often choosing from a list of options for materials, size and shape), and estimate the cost for different parts of the project (such as steps, landings or balustrades) per unit or metre.

Included in the estimations for each separate part of the project are those for the number of engineering hours. TrapCalc then multiplies these hours by a constant hourly rate and sums them up, summarising the engineering costs for the entire project. Importantly, in interviews, the calculators indicated that after filling in a preliminary cost per part of the staircase, they look at the total number of engineering hours allocated for the project to see whether it looks realistic.

The calculators estimate design time based on assumptions about the size and complexity of the project, discounting for repetitive work. They repeat this process for each part of the staircase, then look at the total number of hours allocated for the project to see whether it seems “good”. If they consider the total estimate too high, they go back to previous parts to in- or decrease the estimates to manipulate the total cost. Importantly, the calculators noted that in this phase they do not purely consider the realistic number of hours spent on engineering, but also the ability to sell the product for the associated cost. As such, the distinction between pricing and cost estimation is unclear. It would not be surprising to find that the average estimated engineering cost is below the actual engineering cost, leading to decreased profit margins in return for a higher chance of obtaining an order.

2.2 Literature review of cost drivers for design & engineering

We conduct a systematic literature review of cost drivers in design and engineering, as a basis for this study’s cost estimation. A detailed description of the search strategy and selection process is documented in Appendix B. Of an initial 121 sources, four papers were selected. All offer several applicable cost drivers for design and engineering at EeStairs.

Table 1 shows the identified cost drivers. Xu and Yan (2006) offer the most detailed breakdown of cost drivers, many of which are supported by the other three papers. Based on their work, we divide the cost drivers into four categories and discuss them below.

Product characteristics
Product characteristics are included in each of the papers. Though authors define these cost drivers differently, most relate to the number of actions to be taken and to the complexity of these actions. As noted by Bashir and Thomson (1999), the definition of product complexity as simply the number of functions to be designed “does not give a good picture of design complexity [because] it assumes that all the functions are equally difficult to develop, which is not true.”

Two academic papers attempt to distil product complexity. Xu and Yan (2006) define product complexity as a combination of the products’ structure, size, and shape. Grabenstetter and Usher (2013) further break down the cost. Functional requirements, according to the authors, are the “specific types of functionality which will provide [an] intended behaviour. To achieve these functionalities, they state that several basic components, or components that are “an intrinsic requirement of most jobs”, are usually combined. The complexity is then further heightened by the interdependencies between those components and functionalities, and the technological complexity thereof. And finally, the authors state that the number of subsystems (i.e. functional parts by which a product may be divided) can give an indication of the design effort.

Design process
Next to the product’s characteristics, the design process also influences the design effort. In this category, Xu and Yan (2006) defined the cost drivers of standardisation, process control, and concurrency. The former two were echoed by Bashir and Thomson (1999), defining it as “use of a formal process,” noting that design time is more predictable if the engineering process is standardised. Salam et al. (2009) also named concurrency, using the definition by Winner et al.

(1988): “Concurrent engineering is a systematic approach to the integrated, concurrent design of products and their related processes, including manufacture and support, ... intended to cause the developers ... to consider all elements of the product life cycle.”

Design team

Several studies indicate the importance of the experience of designers for completing the project successfully (Salam et al., 2009). This is because experienced designers are more adept at handling complex information, spend less time thinking about the physics, and easily come up with a multitude of solutions compared to inexperienced designers (Bashir and Thomson, 1999). Efficient communication and collaboration is another important cost driver. Bashir and Thomson (1999) found that over 35% of the total design effort is spent on direct communication, underscoring the impact of team size and communication efficiency on the design effort.

Design conditions

Finally, we will discuss several cost drivers that do not fit into a category per se. One driver is the availability of data (on previous projects), with Grabenstetter and Usher (2013) noting that all the firms they observed attempted, as early in the process as possible, to find similar past jobs which could be used to quote, design and build a new job. Another such cost driver is the presence of regulations and standards, which play a big role within construction and differ significantly between countries. The more strict the regulations, Grabenstetter and Usher (2013) note, the more difficult and expensive a project is.

Categories	Xu and Yan (2006)	Bashir and Thomson (1999)	Salam et al. (2009)	Grabenstetter and Usher (2013)
Product characteristics	structure	product complexity	type of design; degree of change	number of functional requirements and basic components
	size			
	shape			
	added demands	technical difficulty: severity of requirements, use of new technology		number of design interdependencies, technologies and sub-assemblies
Design process	standardisation	use of a formal process		
	process control			
	concurrency		concurrency	
Design team	collaboration	management complexity: team size, methods of communication		
	individual experience	experience, skill and attitude of team members	experience of personnel	
	individual skill			
	dedicated spirit			
Design conditions	design tools	use of design assisted tools		
	management support			
	available data			presence of a reference job
				number of regulations and standards

Table 1: Design & engineering cost drivers found in the systematic literature review.

2.3 Literature review on estimating the design & engineering costs

In this section we review literature on variables that can estimate engineering costs, and the challenges involved.

Estimating design cost

Benedetto et al. (2018) conduct extensive interviews with designers, identifying the following aspects as contributing positively to the design effort quotation:

- Knowledge, or skills required to develop a project's quotation. The authors divide it into two distinct types: explicit knowledge and tacit knowledge.
- Execution, which is related to a professional's knowledge level, in particular the tacit knowledge gained through experience with the subject.
- Design method, because a well-defined design method helps those who estimate the design costs to estimate the time needed for each activity.
- Planning and control, which adds empiricism to the process, allowing evaluation of past projects and their estimations to improve future estimations.

Additionally, an extensive review of systems for machining price quotation found that experienced personnel are essential to determine the subjective factors that influence the generation of a quotation (García-Crespo et al., 2009). The authors suggest VMCs need a knowledge representation model representing:

- Expert knowledge
- Knowledge of other applications
- Detailed estimation models that complement the expert knowledge

Challenges in estimating design effort

Studies identify several challenges to estimating the design effort. Bernardes et al. (2019) note that time estimation poses a unique challenge in design because the reference data available in other fields is not easily accessible. This is corroborated by the fact that design projects are often unique (Kumar, 2008; Rittel & Webber, 1973). Another challenge of finding accurate time estimates is that the activities within the design process are not independent of each other, making it difficult to estimate the time dedicated to each task (Hellebrand et al., 2010). Estimating the design effort is a complex effort, including many uncertainties, thus requiring the combination of various estimation methods (Garcia-Crespo et al., 2009).

Although the factors found in the previous section may be good estimators of design effort in theory, not all these variables are documented in practice. Therefore, Niazi et al. (2006) argue qualitative cost estimation techniques are more appropriate early in the design cycle than quantitative cost estimation techniques - noting that a combination of the two may help provide useful cost estimates at various phases of design and development.

3 Methodology

In this chapter, we discuss our approach (model development) and rationale. The method we propose is adapted from Matel et al. (2019) and consists of three overarching phases: data collection (Section 3.1), model development (Section 3.2) and finally a presentation and discussion of the results (Chapters 4 and 5). The method is visualised in Figure 4.

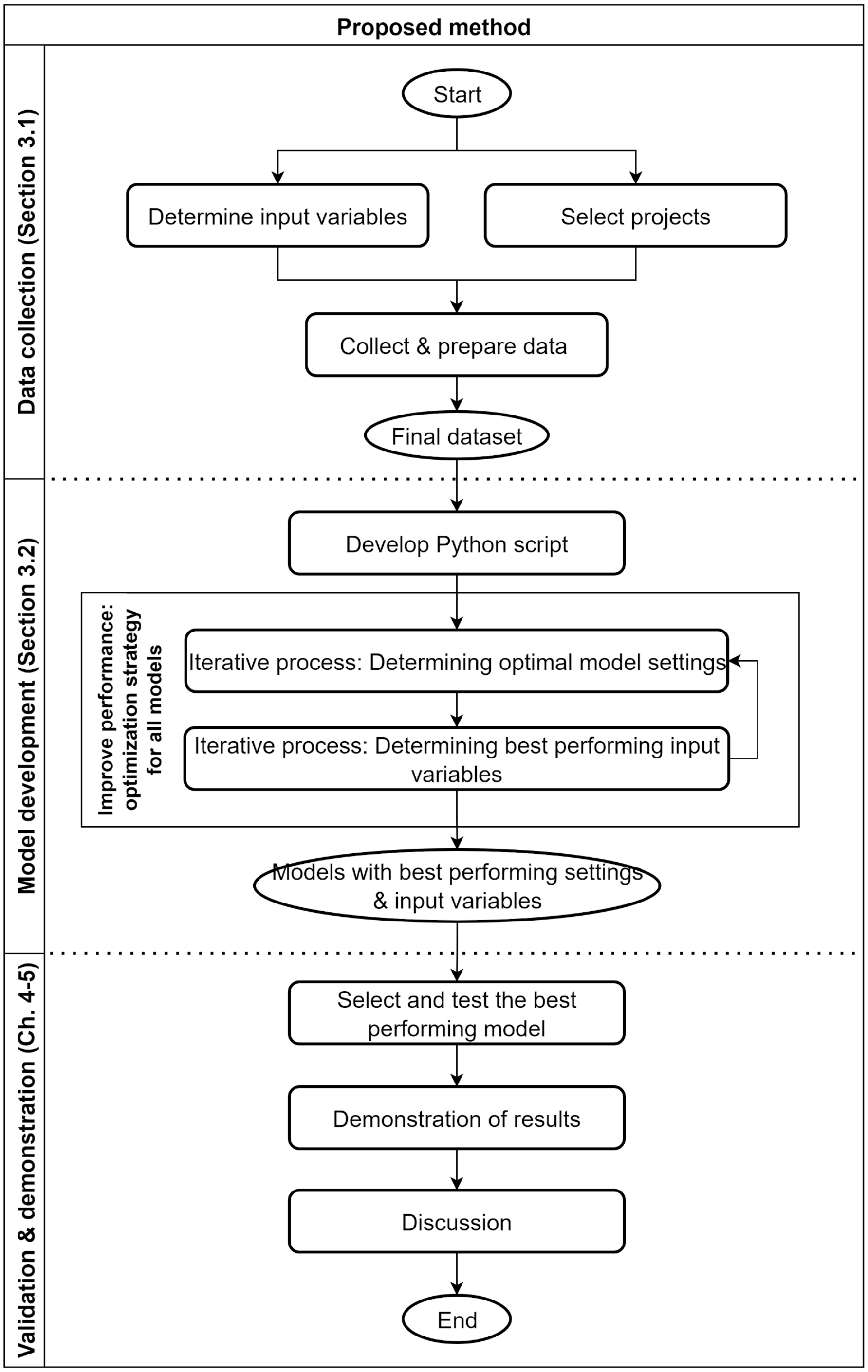

Proposed method
Data collection (Section 3.1)
Start
Determine input variables
Select projects
Collect & prepare data
Final dataset
Model development (Section 3.2)
Develop Python script
Improve performance: optimization strategy for all models
Iterative process: Determining optimal model settings
Iterative process: Determining best performing input variables
Models with best performing settings & input variables
Validation & demonstration (Ch. 4-5)
Select and test the best performing model
Demonstration of results
Discussion
End

Figure 4: Proposed method.

We determine the input variables, or the features we will collect data on, in Section 3.1. Following, we decide which projects to in- and exclude in our dataset. Finally, we perform several feature engineering processing steps to make our dataset usable for model development.

Section 3.2 describes our machine learning model. We develop our model with Scikit-learn, a commonly used Python library for implementing machine learning algorithms. It is then optimised by first choosing the best performing training algorithm out of a selection that fits our project, and then by optimising the selection of input variables for the algorithm. The result is a model whose input is variables relating to one of EeStairs' projects, returning an estimate of the number of engineering hours.

3.1 Data collection

In this section, we describe the data collection process. An overview of our data collection and selection process is visualised in Figure 5, with the coloured boxes corresponding to the Subsections 3.1.1 to 3.1.4 where they will be explained.

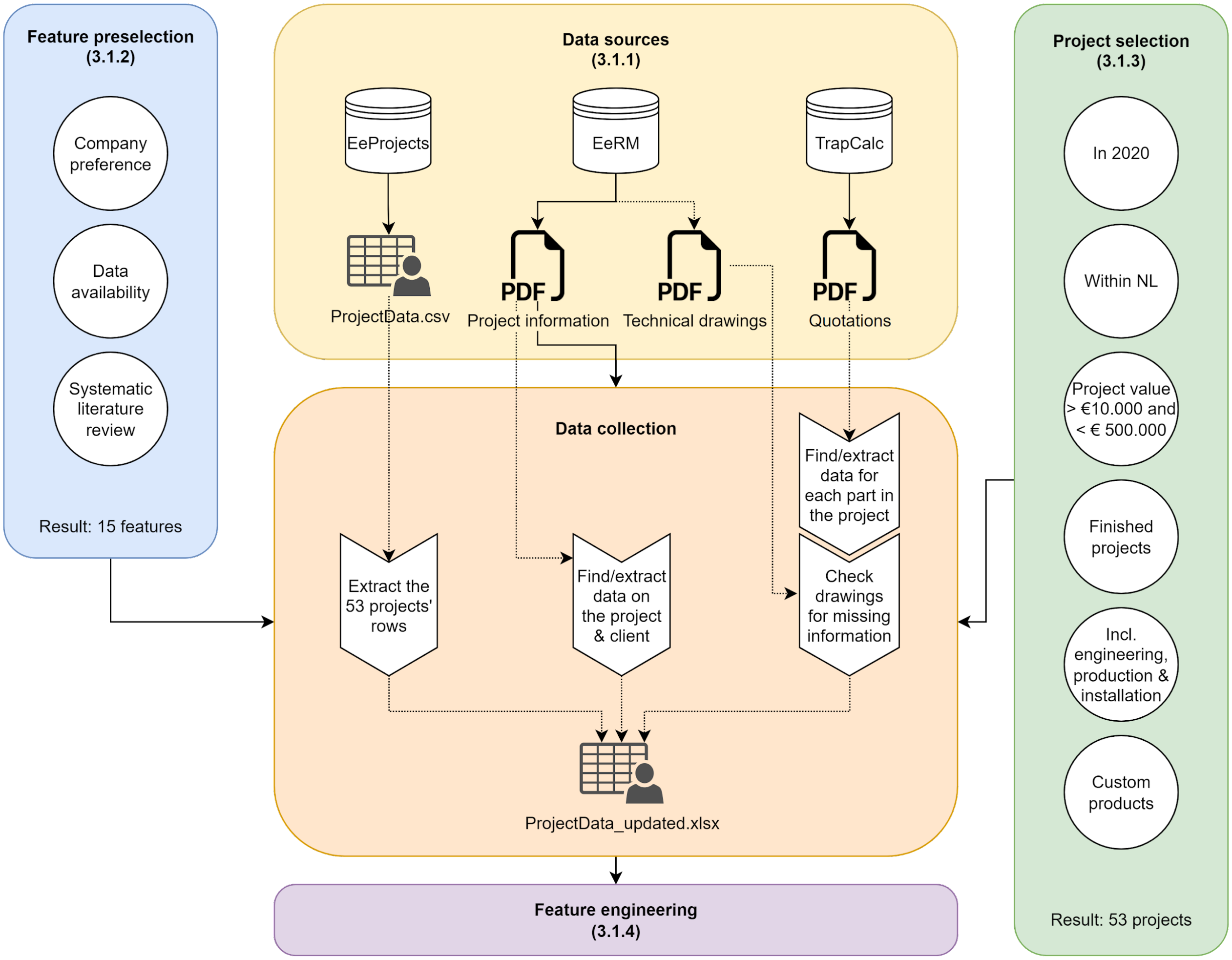

Feature preselection (3.1.2)
Company preference
Data availability
Systematic literature review
Result: 15 features
Data sources (3.1.1)
EeProjects
EeRM
TrapCalc
ProjectData.csv
PDF
Project information
PDF
Technical drawings
PDF
Quotations
Data collection
Extract the 53 projects' rows
Find/extract data on the project & client
Find/extract data for each part in the project
Check drawings for missing information
ProjectData_updated.xlsx
Feature engineering (3.1.4)
Project selection (3.1.3)
In 2020
Within NL
Project value > €10.000 and < € 500.000
Finished projects
Incl. engineering, production & installation
Custom products
Result: 53 projects

Figure 5: An illustration of the data collection process.

3.1.1 Data sources

Figure 6 gives an overview of the data sources within EeStairs. We created a dataset by gathering data from three sources: The Customer Relationship Management (CRM) software called EeRM, a project management service software called EeProjects, and EeStairs' current calculation tool named TrapCalc. What follows is a brief description of these three sources.

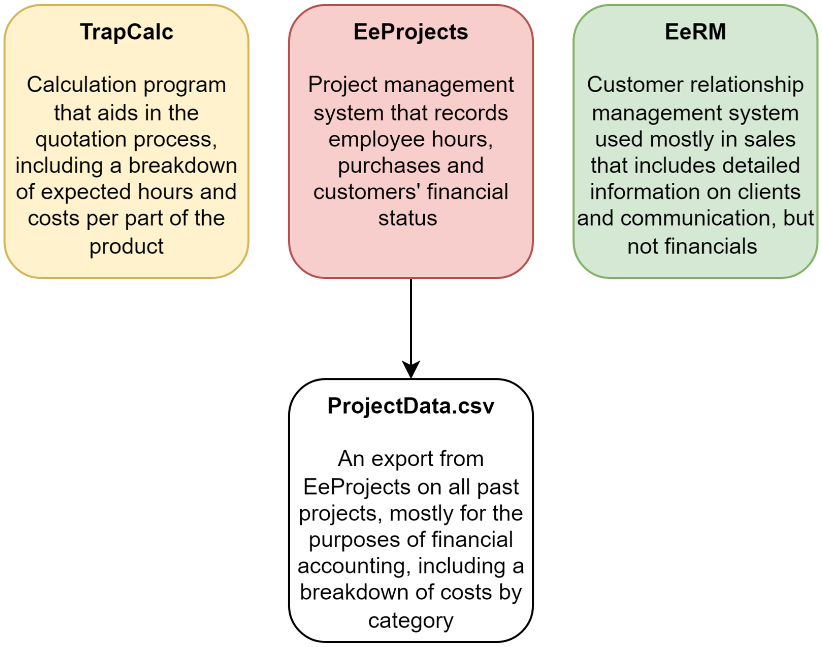

TrapCalc
Calculation program that aids in the quotation process, including a breakdown of expected hours and costs per part of the product
EeProjects
Project management system that records employee hours, purchases and customers' financial status
EeRM
Customer relationship management system used mostly in sales that includes detailed information on clients and communication, but not financials
ProjectData.csv
An export from EeProjects on all past projects, mostly for the purposes of financial accounting, including a breakdown of costs by category

Figure 6: Overview of data sources.

TrapCalc

TrapCalc is the company's quotation software. The company's calculators use TrapCalc to enter all sizes, materials, treatments and other aspects of a project, divided by products (e.g. Stair 1, Balustrade 1 and 2), which are further divided by part (e.g. railing, steps), in order to predict the cost of a project. The program then prints out a quote to be sent to the client, which in this research is our most reliable source to determine the properties of a project and its parts. An extensive description of the program can be found in Appendix A.

EeRM

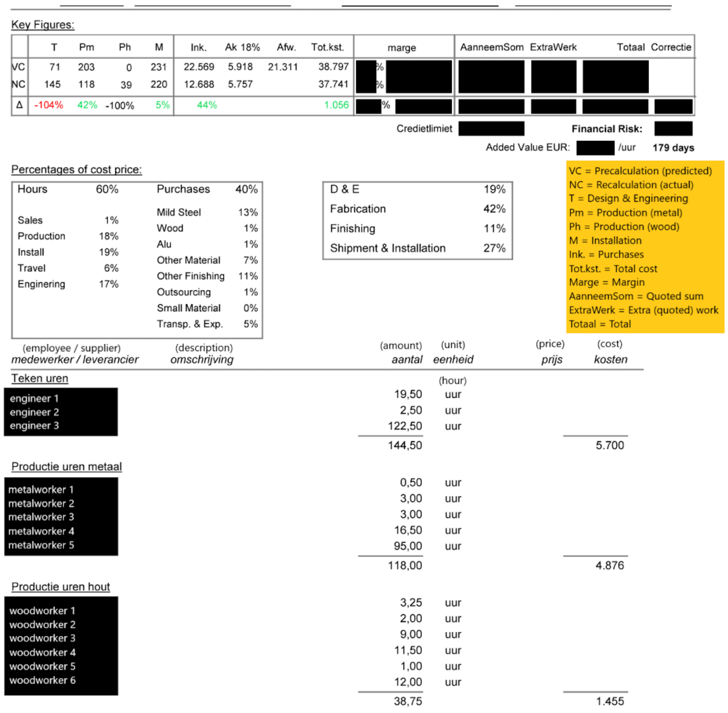
EeStairs has its own CRM software, EeRM. This data consists mainly of high-level quantitative breakdowns (e.g. profit margins), and archives (e.g. e-mail exchanges, quotes, technical drawings, photos of the finished product), but also has some information identifying the type of client. The archived pictures and drawings are used to either corroborate, clarify, or add to the information found in TrapCalc.

EeProjects

EeProjects is the company's project management software. It offers an aggregate of data from TrapCalc, such as the expected total number of engineering hours, and an overview of all hours worked on a project. For each of their projects, EeStairs maintains the following information:

- Project information: the name, code, start- and end date, market segment.
- Client information: the name and address of clients.
- Project management: the project manager, salesperson, quoted price, payables.
- Hours spent: hours worked per person per day, by activity.
- Purchases: material, transportation and outsourcing costs.

Figure 7 shows an example of an overview created with EeProjects. However, because of limited bandwidth from the company's data engineer, not all the data seen in the figure could be exported. As such, in this study, we only use the 'key figures' and 'percentages of cost price' (underlined in the picture) and exclude the cost per individual engineer/employee or purchase. This data has been made readily available through an Excel export (ProjectData.csv), with rows for each project and their respective key figures and cost price percentages.



Key Figures:

	T	Pm	Ph	M	Ink.	Ak 18%	Afw.	Tot.kst.	marge	AanneemSom	ExtraWerk	Totaal	Correctie
VC	71	203	0	231	22.569	5.918	21.311	38.797	[redacted] % [redacted]	[redacted]	[redacted]	[redacted]	
NC	145	118	39	220	12.688	5.757		37.741	%				
Δ	-104%	42%	-100%	5%	44%			1.056	[redacted] % [redacted]	[redacted]	[redacted]	[redacted]	[redacted]

Credietlimiet [redacted] **Financial Risk:** [redacted]

Added Value EUR: [redacted] /uur **179 days**

Percentages of cost price:

Hours	60%	Purchases	40%
		Mild Steel	13%
Sales	1%	Wood	1%
Production	18%	Alu	1%
Install	19%	Other Material	7%
Travel	6%	Other Finishing	11%
Enginering	17%	Outsourcing	1%
		Small Material	0%
		Transp. & Exp.	5%

D & E	19%
Fabrication	42%
Finishing	11%
Shipment & Installation	27%

VC = Precalculation (predicted)
NC = Recalculation (actual)
T = Design & Engineering
Pm = Production (metal)
Ph = Production (wood)
M = Installation
Ink. = Purchases
Tot.kst. = Total cost
Marge = Margin
AanneemSom = Quoted sum
ExtraWerk = Extra (quoted) work
Totaal = Total

(employee / supplier)	(description)	(amount)	(unit)	(price)	(cost)
medewerker / leverancier	*omschrijving*	*aantal*	*eenheid*	*prijs*	*kosten*
Teken uren			(hour)		
engineer 1		19,50	uur		
engineer 2		2,50	uur		
engineer 3		122,50	uur		
		144,50			5.700
Productie uren metaal					
metalworker 1		0,50	uur		
metalworker 2		3,00	uur		
metalworker 3		3,00	uur		
metalworker 4		16,50	uur		
metalworker 5		95,00	uur		
		118,00			4.876
Productie uren hout					
woodworker 1		3,25	uur		
woodworker 2		2,00	uur		
woodworker 3		9,00	uur		
woodworker 4		11,50	uur		
woodworker 5		1,00	uur		
woodworker 6		12,00	uur		
		38,75			1.455

Figure 7: a detailed report on the costs relating to a project. Black boxes censor names and other sensitive data, and the yellow box includes English translations for all (necessary) Dutch words or abbreviations.

3.1.2 Feature selection

We have to aggregate and process some data the company provides before we can start the data analysis. Because this can take a lot of time, we first make a selection of features that we expect will be informative for predicting the number of engineering hours.

In Appendix C, we select variables based on a combination of the cost factors found in the literature review in Section 2.3, and those which the company has indicated to be preferred and/or useful features. After combining the two, we inspect these variables to see which features we can include in practice and arrive at a final feature set which we present in this section.

In order to customise the factors to EeStairs' goals and constraints, we divide the product-related factors into two categories: staircases and balustrades, as these two are distinctive products that should be operationalised differently. Additionally, we introduce those project-wide factors that relate to the client & the project.

Client & Project
These are the factors relevant to all balustrades and staircases that are part of the project, or relevant only to the client. In Table 2, we outline the chosen factors, their variables, and our reasoning as it applies to the choice of factor and/or operationalisation.

Factor	**Variables**	**Reasoning**
Experience & type of collaboration	client: {private, contractor, designer} sector: {residential, office, retail, public, (semi) public}	Information on, for example, the exact size of the client's team cannot be found in the data. The type of client and the sector they operate in are the closest variables we could find.
Travel distance	address	The travel distance formed a part of EeStairs' factors regarding site surveys and design meetings, and the address is easily obtained.
Presence of a BIM	{yes, no}	If the client provided a BIM (Building Information Model), it should be findable in EeRM's file archive. Presence of a BIM should remove some work from an engineers' plate.

Table 2: Client & project features.

Staircase
For each staircase, we have the features shown in Table 3.

Factor	**Variables**	**Reasoning**
Size	width: millimetres height: millimetres	These are four variables commonly available from

	riser: millimetres tread: millimetres	EeStairs' internal data, referring to the size of the staircase.
Shape	shape: {straight, winding}	This information is found in drawings and quotes. While EeStairs uses more categories, we decided to limit it to these two: the laboriousness factor will account for further complexities with regard to shape.
Repetition	0%: Everything has to be designed from scratch. 20: A small amount of this part can be copied from another project or part. 40%: A larger amount of this part can be copied from another project or part. 60%: This part is a copy of another project or part, but some major changes need to be made (i.e. the radius of the staircase, or there is an addition such as a landing). 80%: This part is a copy of another project or part, but some simple things need to be changed (i.e. the staircase is a bit longer or wider than the other, there are some corners or stops in the balustrade). 100%: This part is a one-on-one copy of an earlier project (such as standard balustrades) or of another part in this project (such as when two of the same stairs are built). In case of doubt between two categories, e.g. 10% or 30% can be used.	For repetition (derived from 'degree of change' and 'presence of a reference job'), we decided on a 0-100% scale: some products are full copies of each other, while others have slight or larger differences.
Laboriousness	1: Not laborious (parts of the staircase are extraordinarily large, simply shaped and similar to each other.) 2: Not too laborious 3: Average 4: Somewhat laborious 5: Very laborious (the staircase is very detailed, shapes are very complex or parts are all very different from each other.)	This factor should account for the additional complexities that are not covered by the previous factors.

Calculation	{no, detail, structural}	Some clients need a structural or detail calculation to be done, which entails additional costs.
Balustrades included	{0, 1, 2 sides of the staircase is included}	In most cases, staircases and their balustrades are calculated separately. Some staircases, however, have an embedded staircase. Here, the balustrade and the beam holding the steps together is created from a single piece. This factor is to account for those cases.

Table 3: Staircase features.

Balustrades

The other important component in EeStairs' projects are balustrades. Table 4 shows the variables that relate to the engineering cost of the balustrades. Generally speaking, there are two types of balustrades: standalone balustrades, and those attached to a staircase. Since the relevant cost influence of this is already covered by the shape factor, however, this difference is neglected for the purposes of this study.

Factor	**Variables**	**Reasoning**
Size	Length: metres	Length is the main factor for balustrade size used by EeStairs.
Parts	Corners: count Parts: count	The number of corners and loose balustrades a specific type of balustrade consists of determines how many of them need to be 'placed' in a drawing, increasing the amount of time.
Shape	Shape: {straight, rising, curved, curved + ascending}	Determines the complexity of the balustrade shape.
Type	Type: {custom-made, handrail, TransParancy, GroovEe, FlatRhythm, Cells}	Determines whether components of the balustrade need to be custom engineered, or have been created before. It also records the difference between a handrail and a balustrade, the latter of which takes more time.

Repetition	Group: letter	Determines which balustrades in a project are part of the same design style which, together with the ‘type’ variable’, helps us to determine how many unique designs have to be made for this project.
Detail	1: No detail (it’s a standard type of balustrade.) 2: Not detailed (i.e. it’s a standard type of balustrade with a non-standard height, or it’s a simple handrail.) 3: Average (a fairly straightforward custom-made balustrade, or a more complex handrail.) 4: Detailed (the balustrade is custom-made, of regular complexity.) 5: Very detailed (the balustrade is custom-made, and includes many complex parts.)	This shows the balustrade's design complexity.

Table 4: Balustrade features.

3.1.3 Project selection

EeStairs has worked on 4779 different projects between the years 2000 and 2021. Due to time constraints and limited data accessibility, we did not use all projects within this time period. Table 5 shows the categories based on which we selected projects from the entire dataset and Table 6 describes the projects we exclude, and for what reason. After this selection process, we are left with a final dataset containing 53 projects.

Criteria	**Reason**
Finished projects	Ongoing projects are excluded as the engineering costs are still changing.
Location within The Netherlands	Projects within The Netherlands are the focus of my supervisor’s research at EeStairs. Projects in other countries introduce several complexities, such as different regulations and/or outsourced design work, which are outside the scope of this thesis.
Consists of custom products	Standardised products such as EeStairs’ 1m2 (a staircase that fits within one square metre) are outside the scope of this project, because we focus on custom-made products.
Includes engineering, production and	We only include projects that comprise all aspects of the production process. With this, we remove some outliers, such as design

installation costs	projects or projects that are significantly outsourced.
Project value over EUR 10.000	We exclude very small projects with this threshold, which removes outliers.
Project value under EUR 500.000	We exclude extraordinarily large projects to remove outliers.
Projects from 2020	We focus on the projects executed in 2020 in order to further reduce the quantity so the data collection fits within our time constraints. We decided that a recent year is the most suitable for our purposes, because throughout the years there have been some changes to both the reporting methods and the work process.

Table 5: Selection criteria for narrowing down the scope of the data.

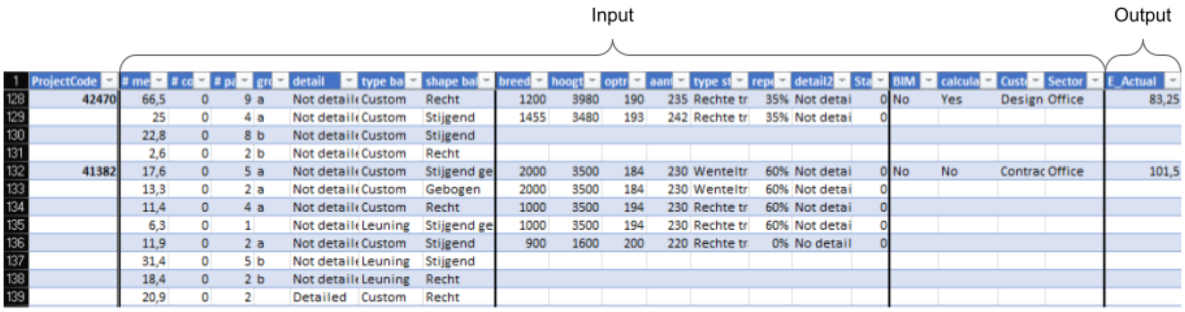
Project	**Reason for exclusion**
43250	This is a maintenance project.
43276	This project includes a significant part (50% of the total price) that is neither a staircase nor a balustrade.
41621	This project almost entirely consists of steel beams and supports for flooring and walls.
40827	Most of this project was initially designed but later scrapped, causing a large discrepancy between the quoted product and the final product.
40769	Like the above, many parts were scrapped and/or added after writing up the quotation. Most of the project was steel construction works rather than stairs or balustrades.
40042	The quotation did not describe the product in sufficient detail for our analysis, and drawings were not included.

Table 6: Excluded projects.

3.1.4 Feature engineering

Having selected a dataset, we can now proceed to prepare the data for analysis. To analyse our data, it must be machine readable, the dimensions of each instance must match, and features should be independent to do a regression analysis (full details in Section 3.2).

The first and primary obstacle is the dimensionality of the data. The features of our collected dataset are descriptive: every project comprises multiple balustrades and staircases (with an average of 7 parts and a maximum of 24). However, the output, the total number of engineering hours spent during the entire project, is a scalar value (see Figure 8). For our analysis, it is necessary to change the data structure so every output has 1 row of input features.



1	ProjectCode	# me	# co	# pa	gro	detail	type ba	shape ba	breed	hoogt	optr	aant	type st	repe	detail2	Sta	BIM	calcula	Custo	Sector	E_Actual
128	42470	66,5	0	9	a	Not detail	Custom	Recht	1200	3980	190	235	Rechte tr	35%	Not detai	0	No	Yes	Design	Office	83,25
129		25	0	4	a	Not detail	Custom	Stijgend	1455	3480	193	242	Rechte tr	35%	Not detai	0					
130		22,8	0	8	b	Not detail	Custom	Stijgend													
131		2,6	0	2	b	Not detail	Custom	Recht													
132	41382	17,6	0	5	a	Not detail	Custom	Stijgend ge	2000	3500	184	230	Wenteltr	60%	Not detai	0	No	No	Contrac	Office	101,5
133		13,3	0	2	a	Not detail	Custom	Gebogen	2000	3500	184	230	Wenteltr	60%	Not detai	0					
134		11,4	0	4	a	Not detail	Custom	Recht	1000	3500	194	230	Rechte tr	60%	Not detai	0					
135		6,3	0	1		Not detail	Leuning	Stijgend ge	1000	3500	194	230	Rechte tr	60%	Not detai	0					
136		11,9	0	2	a	Not detail	Custom	Stijgend	900	1600	200	220	Rechte tr	0%	No detail	0					
137		31,4	0	5	b	Not detail	Leuning	Stijgend													
138		18,4	0	2	b	Not detail	Leuning	Recht													
139		20,9	0	2		Detailed	Custom	Recht													

Figure 8: The input and output data for 2 projects.

To do so, we are presented with three options:

1. **Increase the quantity of output data points.** One option is to create a formula that distributes engineering hours over all different parts (e.g. find 6 and 13 output variables for projects in Figure 8).
2. **Decrease the number of input data points**. Another option is to merge input features through, for example, addition, averaging or combining features in more complex ways.
3. **Increase the number of features.** Accommodate for the largest project in the data set while setting empty values to 0 for smaller projects. For the ‘metres’ feature, for instance, this would result in a ‘Metre1’ feature for the first balustrade in a project, ‘Metre2’ for the second, etc.

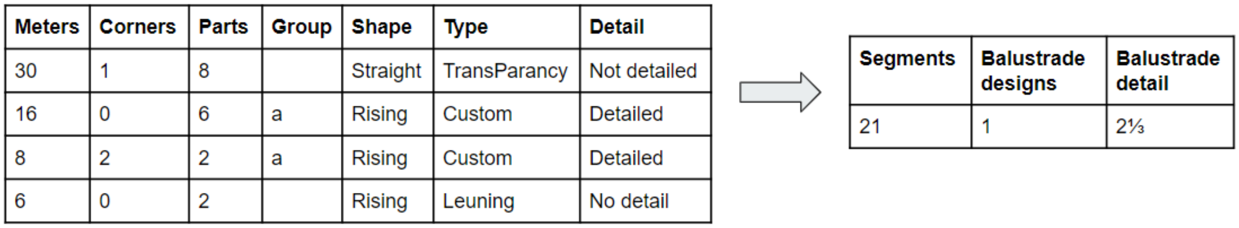
We choose option two. The third option appears undesirable because the result would be a data set with more columns than rows, which is nearly unusable for machine learning. Option one relies on splitting up the output value (and other overarching project-related features) over several parts in a way that is, at best, arbitrary. While it would increase the amount of training data from 53 (projects) to 368 (parts), it is uncertain whether this weighs up to the information loss due to the detachment from the projects a product belongs to. The second option appears most favourable as we can minimise the information loss by carefully engineering the features.

To condense the data for each project into a single row while minimising the information loss, we utilised different methods. What follows is a brief description of the condensation process for each category of feature: balustrade, staircase, and other.

Balustrade

This category comprises quantitative features (metres, corners, parts) that are highly correlated. We choose to reduce them to the ‘segments’ feature, adding the number of ‘corners’ and ‘parts’ together. These appear more meaningful than the number of metres because extending a balustrade is (in terms of design) a relatively simple operation. Having to place parts in different places, and making a balustrade run across a corner, is therefore a better indicator of design effort. While the shape of the balustrade also impacts the design effort, we choose not to split the ‘segments’ among the four different shapes, losing the ‘shape’ feature.

The other qualitative features, type, group and detail, relate more to the time to create the balustrade design, which is independent of how often this design is copied. We reduced these to two features: the number of balustrade designs and the average detail of those designs. Figure 9 demonstrates our operations.



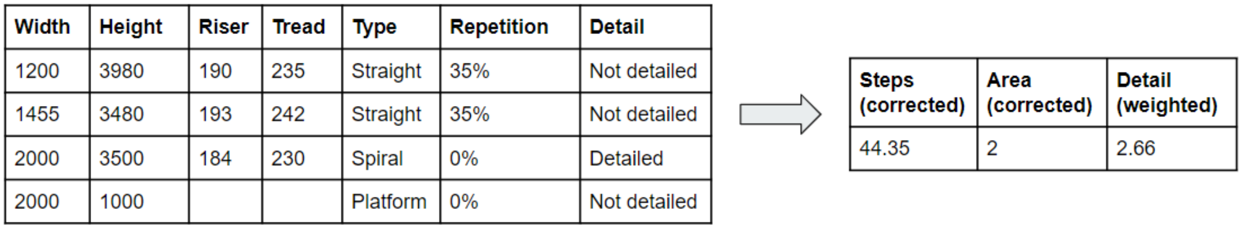
Meters	Corners	Parts	Group	Shape	Type	Detail
30	1	8		Straight	TransParancy	Not detailed
16	0	6	a	Rising	Custom	Detailed
8	2	2	a	Rising	Custom	Detailed
6	0	2		Rising	Leuning	No detail

Segments	Balustrade designs	Balustrade detail
21	1	2⅓

Figure 9: An example of how the balustrade features were engineered.

Staircase

It is much more complex to reduce the number of rows for staircases. Our approach is to compute the number of steps for each staircase by dividing the height by the riser (for the first staircase in Figure 10, this is 3980/190 = ~21). For staircases that have platforms, we include the area as well. We correct the number of steps by the repetition quantity, the percentage of the staircase “copied” from other staircase designs within the project or in other projects. Next, we subtract this number from the total sum, which in the example of Figure 10 is formulated as steps = 21*0.65 + 18*0.65 + 19*1 = 44.35. Detail is integer encoded (‘No detail’ = 0, ‘Not detailed’ = 1, … ‘Very detailed = 4’). Finally, we compute a weighted average.



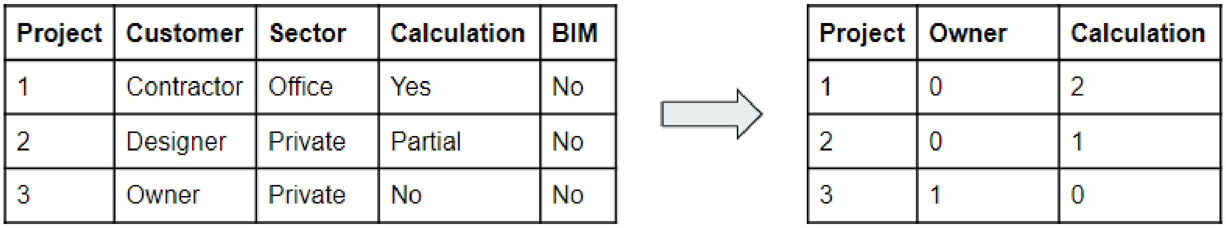
Width	Height	Riser	Tread	Type	Repetition	Detail
1200	3980	190	235	Straight	35%	Not detailed
1455	3480	193	242	Straight	35%	Not detailed
2000	3500	184	230	Spiral	0%	Detailed
2000	1000			Platform	0%	Not detailed

Steps (corrected)	Area (corrected)	Detail (weighted)
44.35	2	2.66

Figure 10: An example of engineering the staircase features.

Other

All other features relate to the entire project, and as such do not need to be condensed. All features in this category, aside from travel time (denoted in hours), are categorical and are integer encoded. Figure 11 shows an example.



Project	Customer	Sector	Calculation	BIM
1	Contractor	Office	Yes	No
2	Designer	Private	Partial	No
3	Owner	Private	No	No

Project	Owner	Calculation
1	0	2
2	0	1
3	1	0

Figure 11: An example of how the other features were engineered.

Finally, ‘sector’ feature is left out because interviews indicated it was viewed as rather inconsequential to the design effort. BIM, conversely, is considered quite impactful but excluded as there is only one known occurrence of a project with BIM, and because company management showed they have low confidence in that data.

3.1.5 Exploratory data analysis

In this section, we provide a brief exploration of the statistics and aspects of the remaining 53 projects. We highlight insights regarding EeStairs’ current estimation accuracy and projects relation to the engineering effort.

Figure 12 shows a bar chart of the 53 projects and their total cost. We observe that the project costs range from €6029 to €381,962, with a median of €25948 and an average of €38817. Figure 13 shows the number of engineering hours spent on each project plotted against the project's cost. It has a linear trendline, that is, engineering hours = 0.00192005 * Total costs + 14.6402, fitted to it. From this figure, we can conclude there is a positive correlation between the number of engineering hours and the total cost of the project, confirming our intuition that larger projects require more engineering work.

Judging from Figure 13, there may be a risk of outliers. Should the two rightmost projects be excluded, the result may be a trend line which is much steeper. However, we choose to keep those projects included: larger projects are already very important for EeStairs and in the future the company aims to further specialise in large and/or complex projects, thus there is a clear need to be able to predict those.

Moreover, the large deviations from the trendline in Figure 13 show there are several other factors that influence the engineering cost besides the project size. This can also be observed in Figure 15, which shows that engineering as a percentage of the project cost varies significantly, with a 50% confidence interval between 0.84 and 4.58.

Figure 14 compares the estimated against the actual number of engineering hours per project. The blue lines (y=x) indicate what a perfect estimation would look like. We can see there is a tendency to underestimate the number of engineering hours. Often, the actual number of hours is more than double the estimate.

Lastly, Figure 16 shows a stacked bar chart for the number of estimated engineering hours per project, divided by those agreed upon at the start of the project (‘Initial’) and those that are added during the project (‘Extra’). Extra hours occur when a client has additional demands after the project was officially agreed upon, for instance because the client requests significant changes or wants a new product.

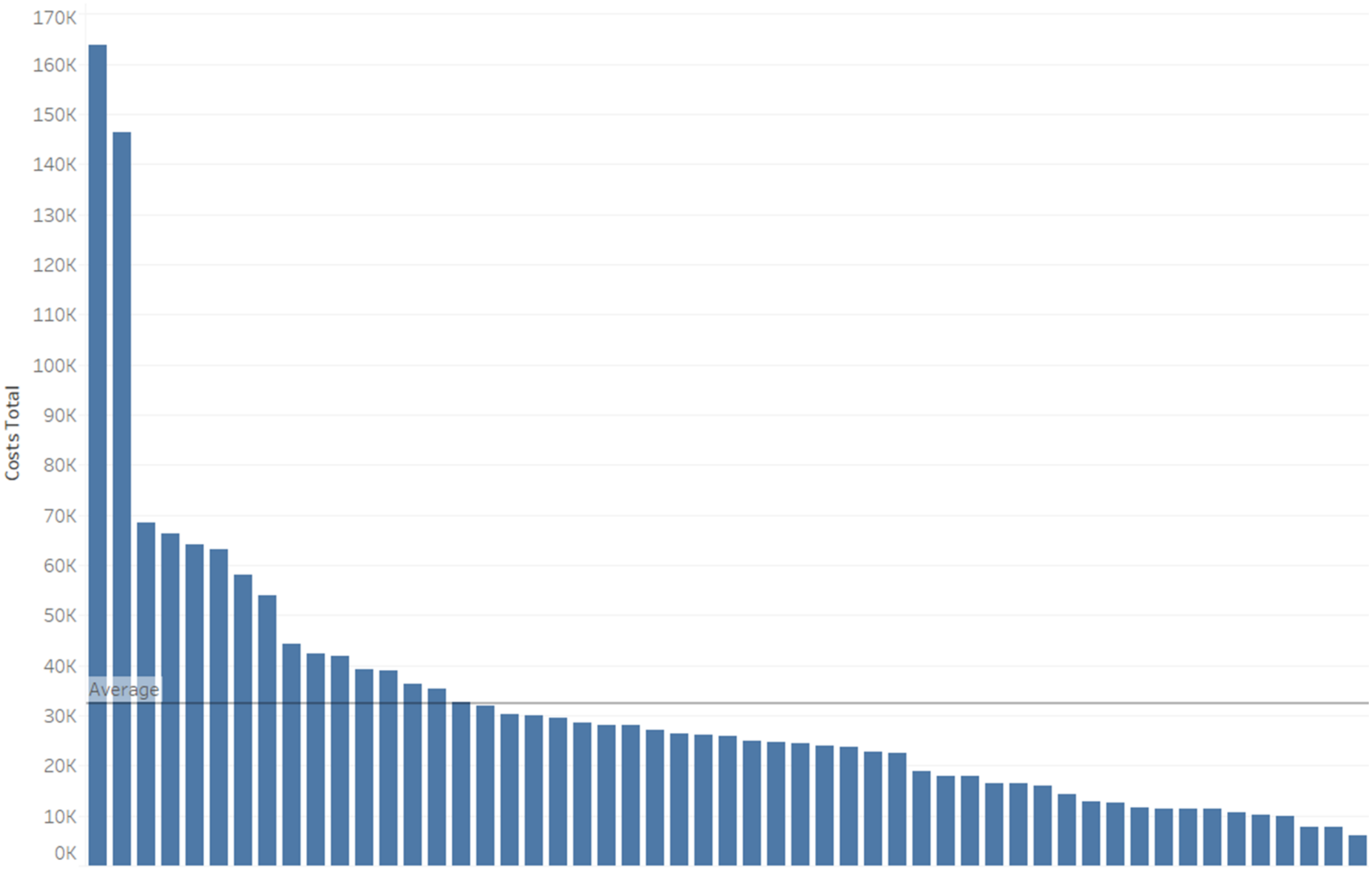
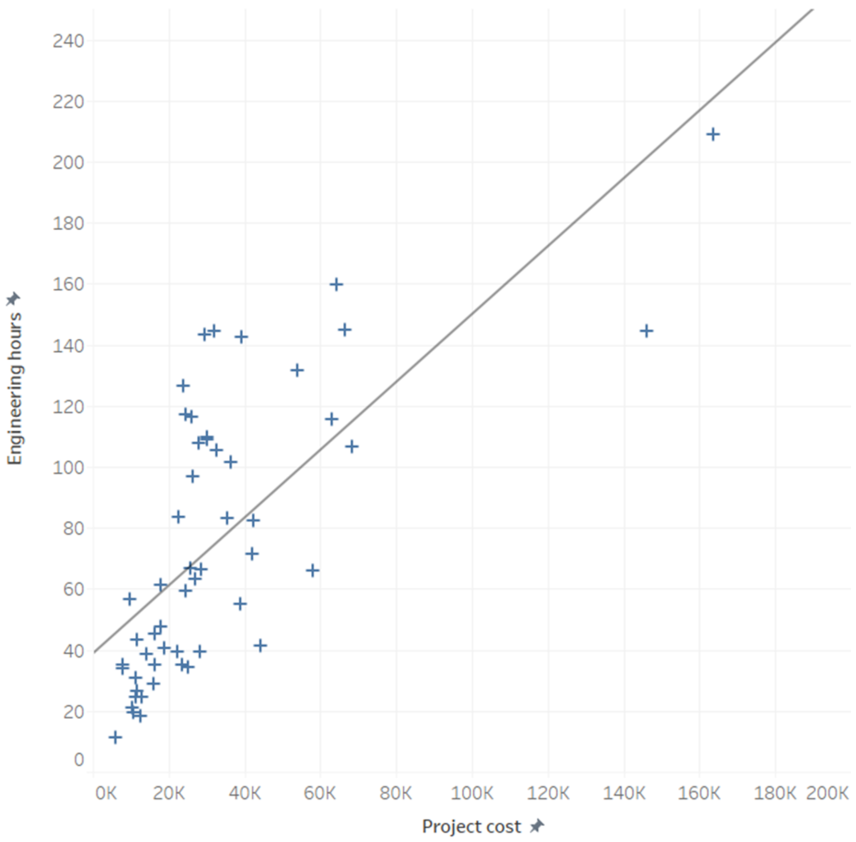
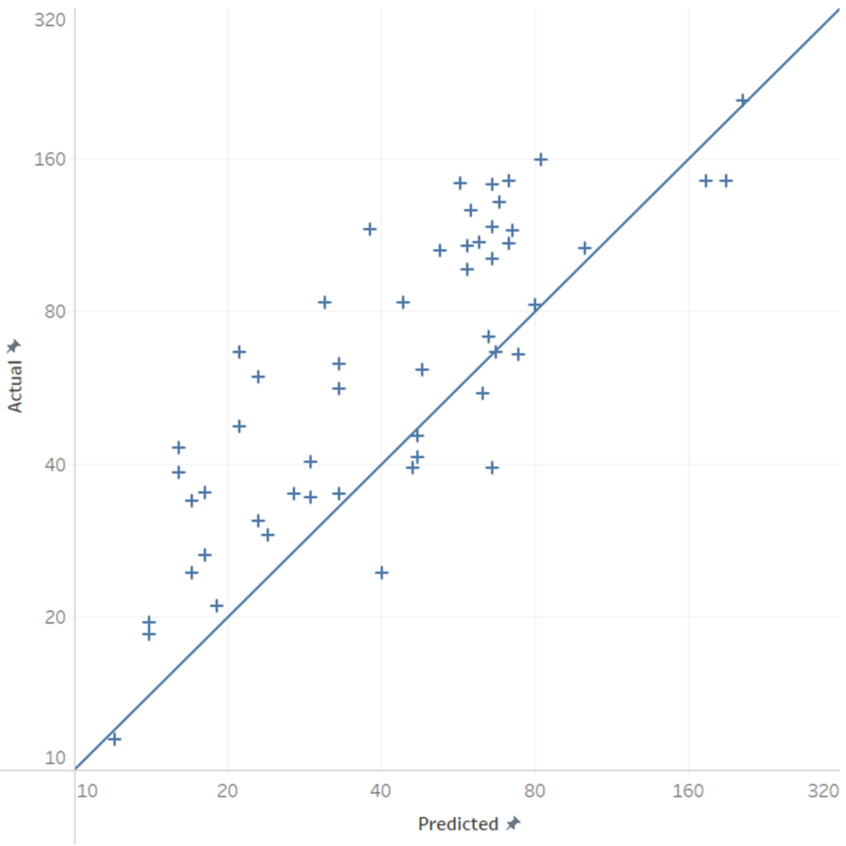


Figure 12: Total costs per project.





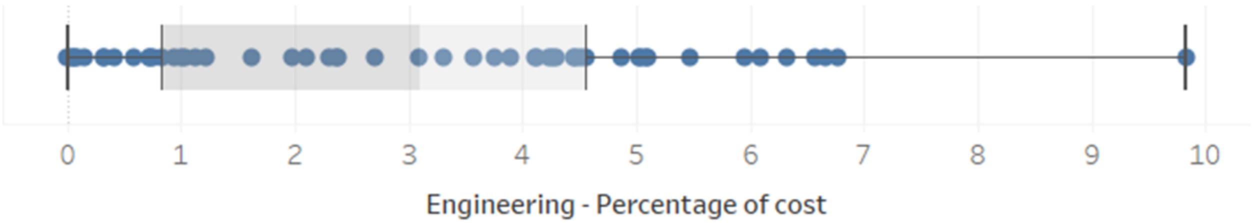


Figure 13 (left): Scatter plot of engineering hours vs. project cost.
Figure 14 (right): Scatter plot of the predicted vs. actual no. of engineering, log. scale. Ideal performance is indicated by a (y=x) line.
Figure 15 (bottom): Box plot for engineering costs as a percentage of the total project cost.

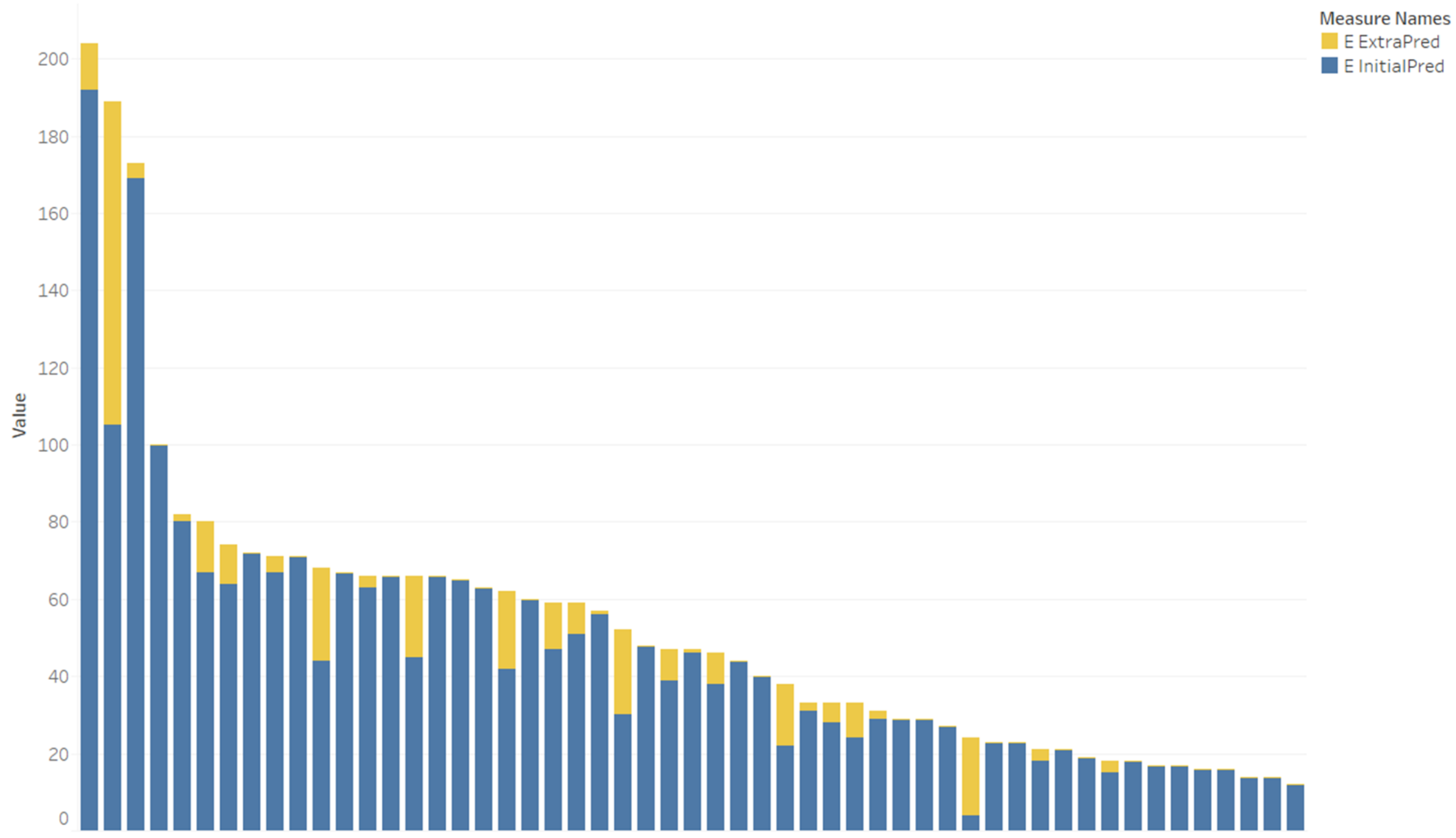


Figure 16: The number of engineering hours per project. In blue those initially agreed upon and included in the original quotation, in yellow the extra hours.

3.2 Model development

We describe our approach to designing and developing our model for predicting EeStairs' engineering costs based on the selected project features. In Section 3.2.1 we introduce regression analysis, and in Section 3.2 the method used to make the most of our data for training and testing. Section 3.2.3 describes how the data was scaled, and Section 3.2.4 introduces the method used to select features to use in the final model. Finally, Section 3.2.5 describes the four regression models which were used in our experiments.

3.2.1 Regression analysis

The statistical method used to develop our model is regression analysis: we aim to predict the number of engineering hours (output variable) based on several features, which contain information about projects from EeStairs. In this section, we explain regression analysis and introduce four variants of regression utilised in this thesis: Multiple-, Ridge-, Bayesian Ridge- and Lasso Regression.

The most common regression model is linear regression. The simplest form, linear regression, uses a single feature (x) to predict an output variable (y). Mathematically, we have the form $y = \beta_0 + \beta_1 x$. Since x and y are known (while training our model), the aim is to estimate β_1 (the regression coefficient) and β_0 (the constant).

Figure 17 shows an example of this process. Blue dots indicate the data points: for example, in the top right we see a point where X = 58 and y = 16. The goal of linear regression is to learn the coefficients such that the error (for instance, the distance from the line to all the points) is minimised. This function can make a new prediction where we only know the value of a feature, but not the value of the output variable. Say X = 0, for example; we find the y-value of the red line at the point where X = 0 and predict y = 5.

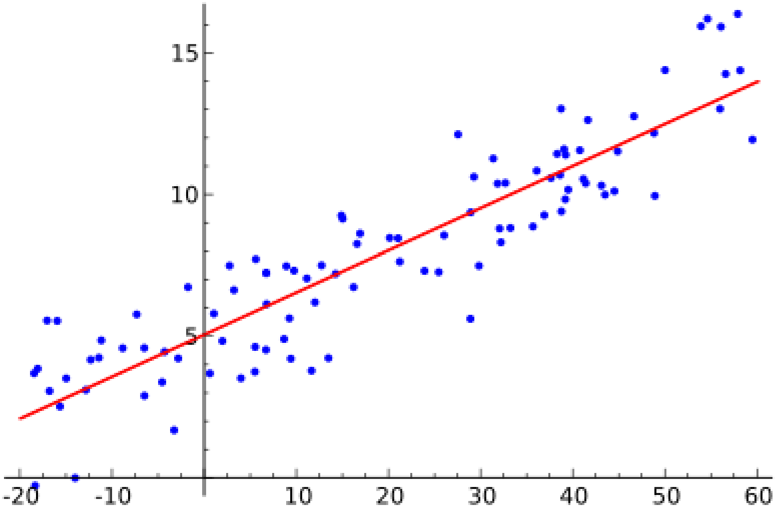


Figure 17: An example of linear regression with one feature (Sewagu, 2010).

As mentioned, regression models estimate the coefficient and constant value to minimise an error, as displayed in Figure 18. This error can be measured in different ways. A common metric is the mean absolute error (MAE), which is the average distance |y - ŷ| from the predicted value (ŷ) to the actual value (y). Another metric is the mean absolute percentage error (MAPE), which defines the absolute error in relation to the actual value - or |y - ŷ|/y. Since these functions are not differentiable, however, the most commonly used metric in machine learning is the mean squared error (MSE), defined as $(y - \hat{y})^2$. In this thesis, we will use MSE to optimise our models while calculating the MAE and MAPE to offer further insight into our results.

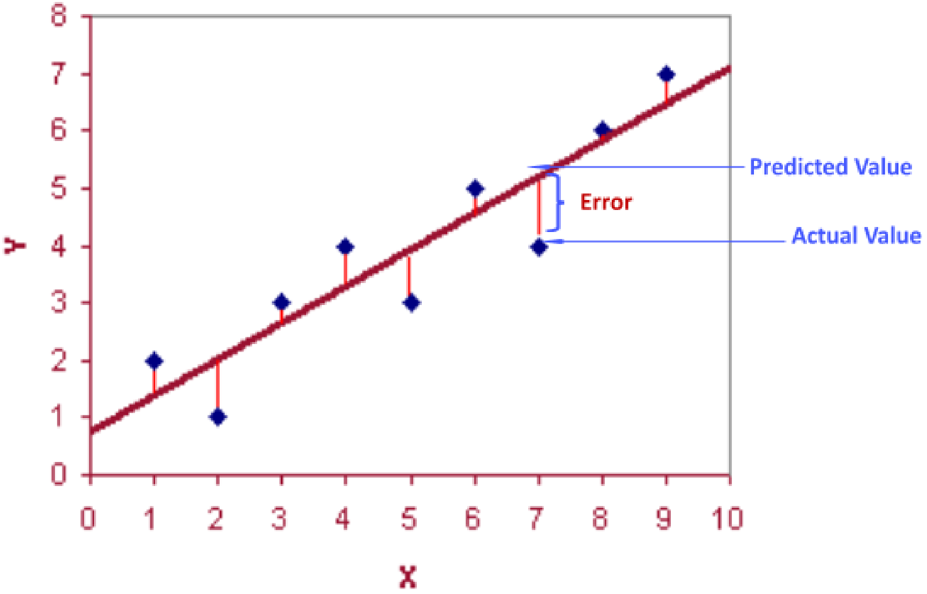


Figure 18: An example of linear regression, illustrating the error (Gupta, 2021).

3.2.2 k-Fold Cross Validation

To correctly assess the model's performance, it is essential to split the data into test and train sets. If our trained model can make accurate predictions on unseen data, the model is capable of generalising from the training set to the test set (Müller & Guido, 2016).

It is possible, however, for the model to overfit (as illustrated in Figure 19), which happens when the model exactly remembers the exact data points, instead of extracting patterns from the data. If an overfit model makes predictions with the test set, we will find its performance to be drastically lower than its performance on the train set.

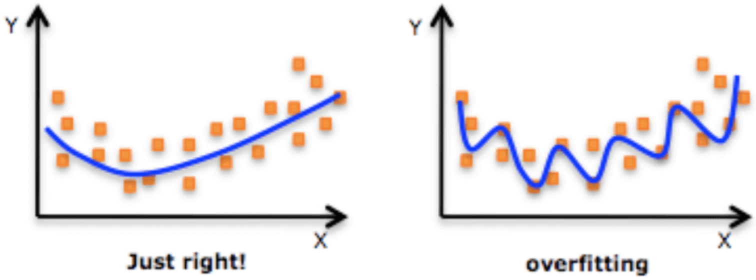


Figure 19: An illustration of an overfitted model (Bronstein, 2017).

Due to the limited size of our dataset, the probability of an imbalanced train-test split is high. The outcomes may significantly vary depending on the selection of projects in the respective sets, and the amount of data available for training could decrease. To mitigate this issue, we utilise k-Fold Cross-Validation, a technique described in Bengio & Grandvalet (2004) and depicted in Figure 20.

In brief, this approach partitions the dataset into k subsets, with typical choices of k being 3, 5, and 10. To maximise the amount of training data, we opted for k = 10 in this study, resulting in each split containing 5 or 6 projects. Then, for *t = 1 to 10*, we use split *t* as the test set and all other splits as the train set. Each of the 10 distinct train-test combinations is a fold. Within each fold, the entire training process is carried out using the training set, and subsequently, the model's performance is evaluated using the corresponding test set.

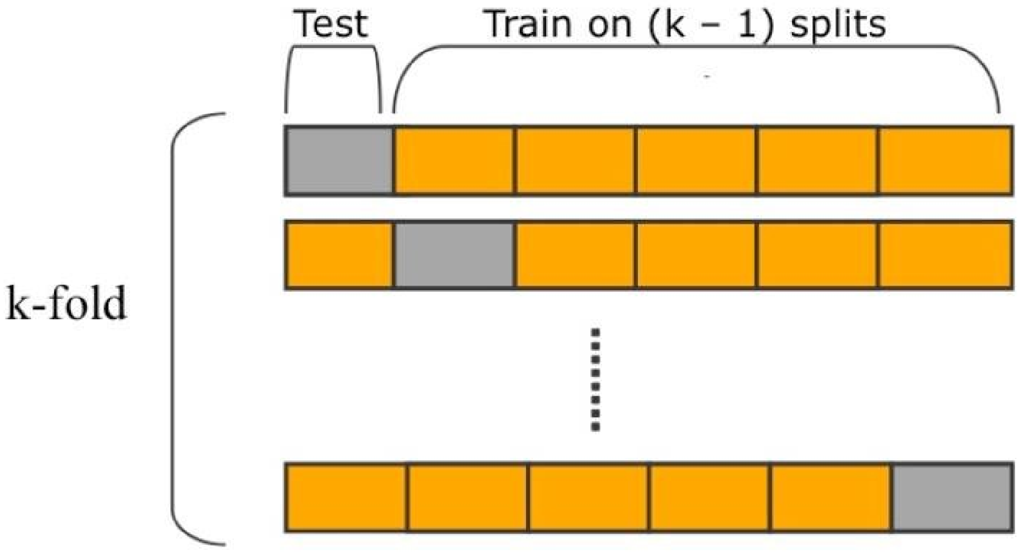


Figure 20: Illustration of k-Fold Cross Validation (Sossi Alaoui et al., 2018).

This process results in ten different scores. The mean and variance over these scores are computed, allowing us to assess the overall performance and to compare it to other models.

3.2.3 Scaling

Scaling is necessary because our current feature set displays a significant variance in data points. For instance, while there are between 0 and 7 balustrade designs across projects, the number of balustrade metres varies from 5.1 to 308.8. As a result, larger values of metres may significantly impact the outcome compared to the number of designs. Scaling guarantees all data is within a comparable range, facilitating parameter comparison. To do so, we apply the StandardScaler method to our features, ensuring each feature's mean is zero and the variance is one, while preserving the data point distance as much as possible.

3.2.4 Feature Selection

To reduce the complexity of our models, a method called SelectFromModel was used (scikit-learn, 2023). This uses the models introduced in the following section to find the optimal number of features to be used to train our model.

3.2.5 Model training

After scaling the training dataset, we train four distinct models: Multiple Regression, Lasso Regression, Ridge Regression, and Bayesian Ridge Regression. In addition, these models were used for feature selection.

Multiple Regression

Since in our project we have multiple features to predict the number of engineering hours, we will use multiple linear regression (scikit-learn, 2023). Instead of only one feature, as illustrated above, it uses multiple (independent) features to predict the output variable. This becomes harder to visualise because an extra dimension is added for every additional feature, but the formula remains familiar: $y = \beta_0 + \sum_{f=0}^{p} \beta_f x_f$, where p is the total number of features.

Lasso Regression

Multiple regression fits a function that suits best to the training data, which does not mean the learned model generalises well. There is a danger of overfitting, where the model performs well on the training data but much worse on data it has not seen before, on testing data. In general, complex models will have a lower bias (or error on the training data) and a higher variance (or error on the testing data): the more complex the model, the more the model's coefficients will change depending on which training data it is offered. As a rule of thumb, one should aim to keep models as simple as possible.

The aim of regularisation is to reduce complexity, by adding a penalty to the features' coefficients to encourage solutions where coefficients are small. This leads to simpler and more interpretable models, which generally perform better on the test data (Müller & Guido, 2016). Lasso Regression, for instance, leverages L1-regularisation, which introduces a penalty term

$\sum_{f=0}^{p} |\beta_f|\lambda$, where λ or lambda determines the magnitude of the penalty (scikit-learn, 2023). This penalty encourages the objective function to minimise the coefficient's magnitude or eliminate features (by reducing the coefficient to 0) that have minimal relevance. The objective function from this approach is: $y = \beta_0 + \sum_{f=0}^{p} \beta_f x_f + |\beta_f|\lambda$

Ridge Regression

Ridge Regression is similar to Lasso regression, using L2-regularisation instead (scikit-learn, 2023). It employs the squared coefficient rather than the absolute value of the coefficient, resulting in a penalty of: $\sum_{f=0}^{p} (\beta_f)^2\lambda$. The most important difference between Lasso and Ridge Regression is that L2-regularisation tends not to entirely eliminate features.

Bayesian Ridge Regression

Bayesian Ridge Regression is a linear model incorporating a Bayesian framework to find a balance between over- and underfitting. The model estimates the distribution of the coefficients instead of a single value, using prior knowledge about the distribution of the coefficients to update the posterior distribution after observing the data. To control the complexity of the model and prevent overfitting, the algorithm adds an L2-regularisation parameter. The model uses Bayes' rule to update the posterior distribution by multiplying the prior distribution by the likelihood function of the data, which it assumes to be normally distributed. The uncertainty of the model is captured by the covariance matrix of the posterior distribution (scikit-learn, 2023; Koehrsen, 2018). Additionally, this model was used because it is thought to work well with small datasets.

4 Results

In this chapter, we outline the results of our model selection, optimisation and training process. The performance of our optimised models and motivation for our final model choice is outlined in Section 4.1. Finally, in Section 4.2, we assess our final models' performance in more detail and compare it to EeStairs' old calculations.

4.1 Optimal model

After several experiments, as described in Appendix D, we ended up with optimised models whose scores are shown in Table 7. Among all the models, Lasso had the best MSE and MAE scores, while Linear and Bayesian Ridge Regression outperformed Lasso on MAPE. Furthermore, Lasso had a considerably lower estimated variance compared to the other models, as depicted in the two columns on the right. Therefore, we selected Lasso as the final model. The MSE scores are also visualised in Figure 21.

Model	**MSE**	**MAE**	**MAPE**	**MSE % increase over train score**	**MSE variance**
LinearRegression	970.51	24.996	0.5350	19.29	1281.85
Ridge (alpha = 8)	941.02	24.737	0.5508	14.68	665.55
BayesianRidge	928.18	24.450	0.5265	16.03	763.66
Lasso (alpha = 5)	888.26	24.179	0.5452	6.77	575.71

Table 7: Several scores and indicators for the four optimised models.

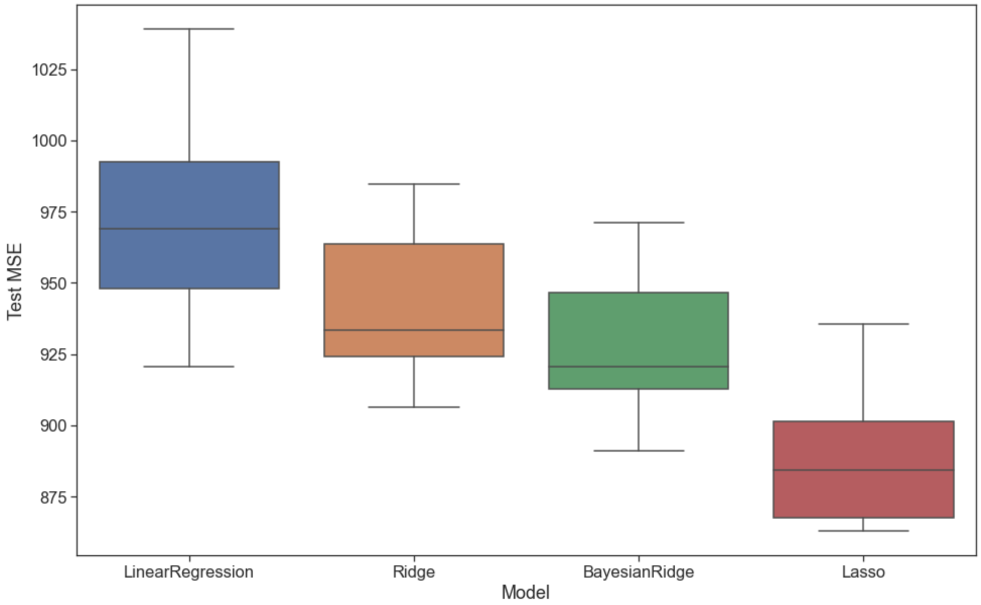


Figure 21: Box plots of the four optimised models' test MSE scores.

4.2 Model performance

Next, we will compare the performance of our final Lasso model with the company's predictions, and analyse its performance. For this, we used leave-one-out (or 53-fold) Cross Validation, generating valid results for all projects within the dataset. All predictions, both by the model and the company, are shown in Figure 22. The resulting error scores, shown in Table 8, demonstrate that the model (which was optimised for MSE) outperforms the company's predictions at the MSE and MAE scores, but has a score significantly worse for MAPE.

	MSE	**MAE**	**MAPE**
Company	1133.36	25.316	0.3483
Model	895.85	24.233	0.5471
Difference	-21.0%	-4.3%	+57.1%

Table 8: Comparison of the model's error scores with those of the company's predictions.

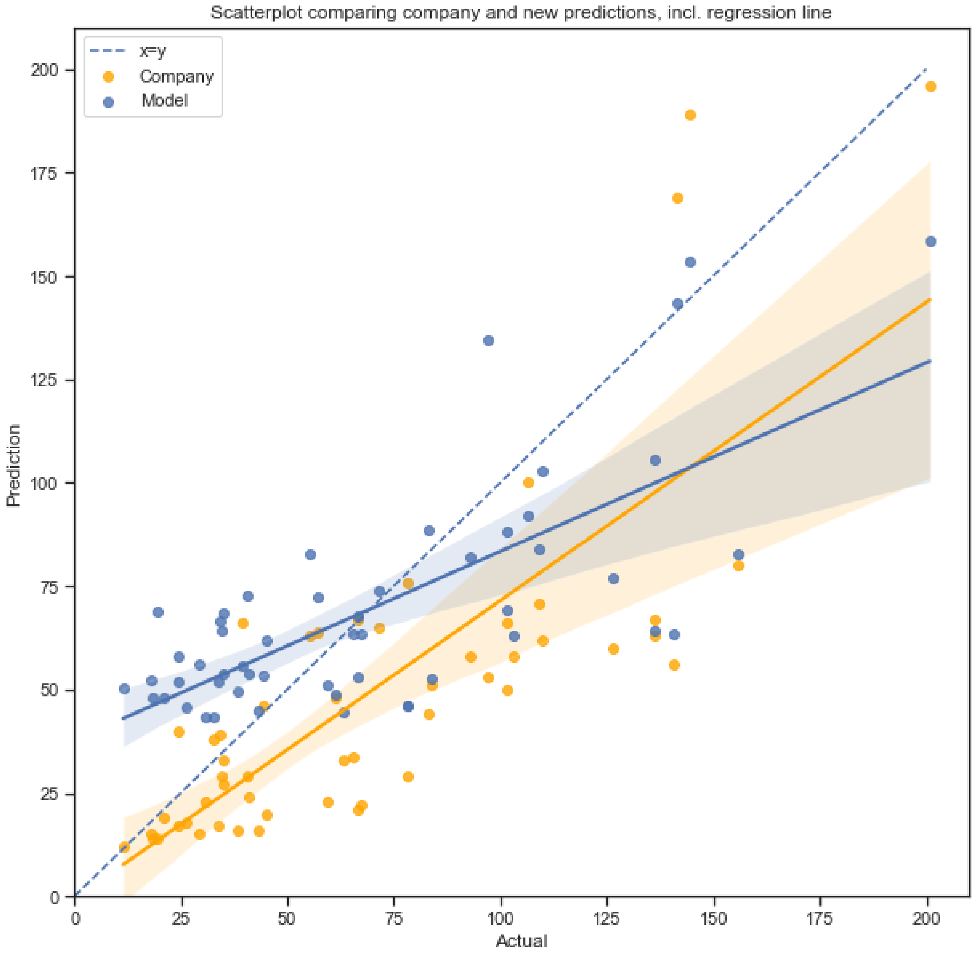


Figure 22: Scatter plot comparing the predicted to the actual number of engineering hours, for our model and the company's, including a trend line for both.

The swarm plots presented in Figure 23 offer detailed observations of the error scores for all individual predictions. The error distributions reveal that the model's errors are centred, with a mean of -0.03, whereas the company's predictions are skewed to underestimate the number of engineering hours, with a mean error of -19.98. Regarding the APE distributions, while the bulk of both methods' predictions have a percentage error ranging between 0 and 75%, our model presents significant outliers of up to 338%.

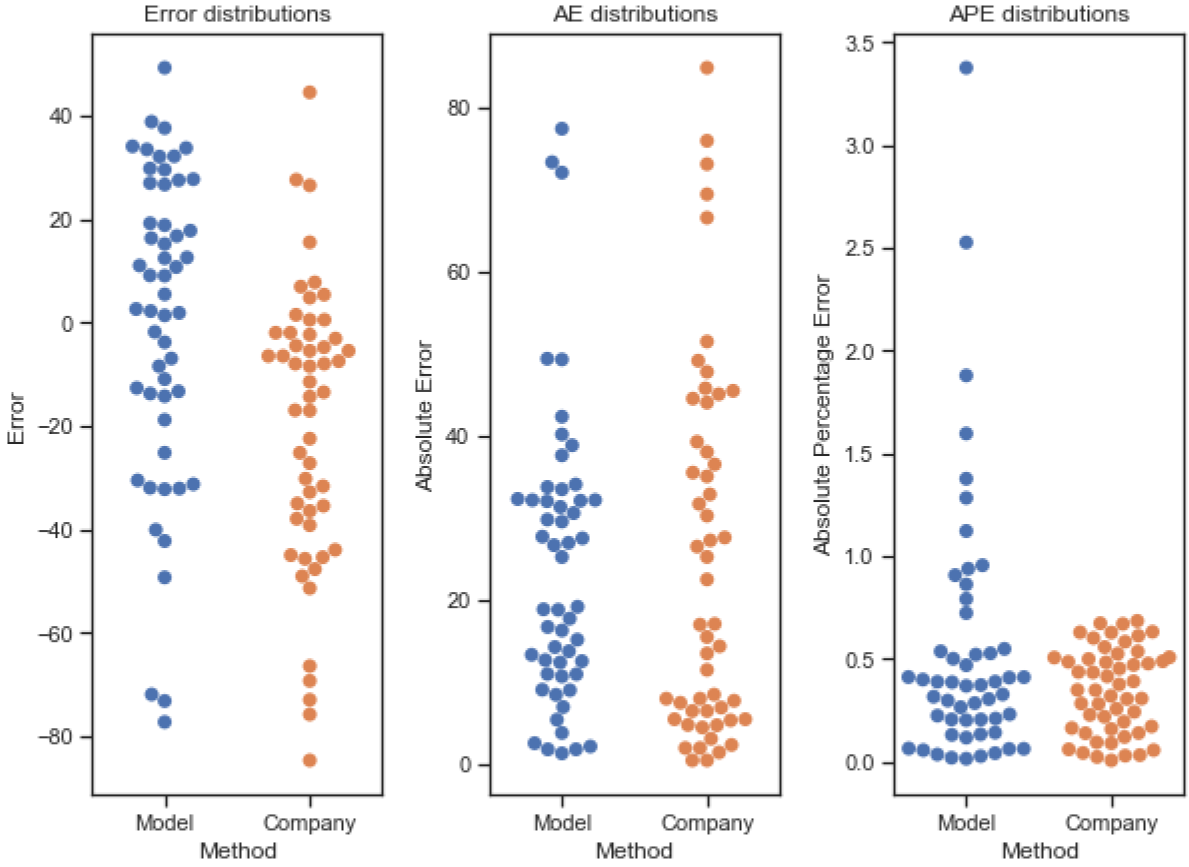


Figure 23: Swamplots of the Error, Absolute Error and Absolute Percentage Errors, for both our model and the company's.

In-depth examination of the APE score reveals smaller projects have the highest percentage errors, as illustrated in Figure 24. We observe a positive Pearson correlation of 0.68. In addition, the models' highest percentage errors are overestimations of actual costs, whereas larger projects are more precise - with a slight tendency to be underestimated.

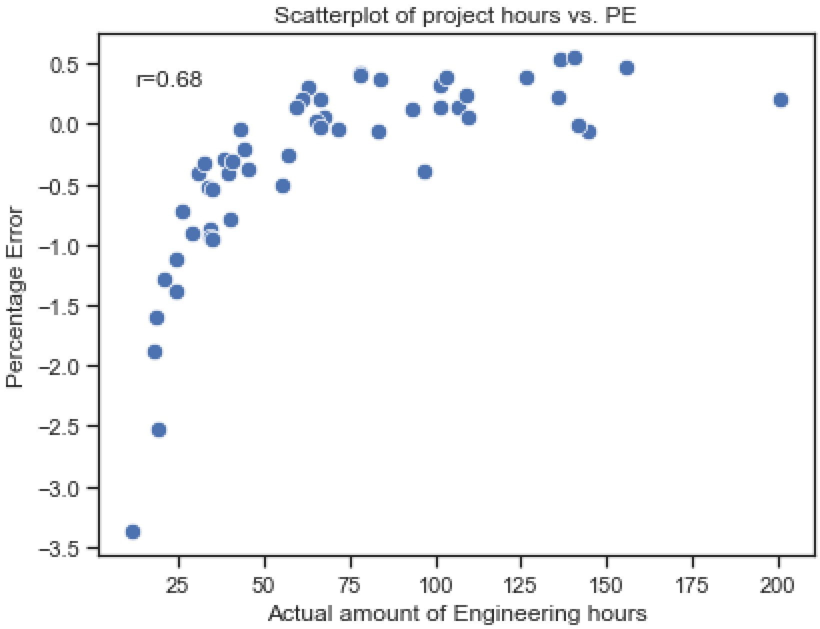


Figure 24: Scatterplot of the percentage error to the number of engineering hours.

In the discussion in the next chapter, we will interpret these findings and their limitations and propose explanations for the observed results..

5 Discussion

In this thesis, we outlined a roadmap for integrating machine learning into EeStairs' quotation process for automation. Our findings show the optimal model outperforms EeStairs' manual quotation calculator as measured by the MSE and MAE. However, its accuracy for small projects is low. In this chapter, we present a thorough evaluation of our model's performance and the limitations of our research, concluding with several recommendations for EeStairs.

5.1 Interpretations

Our model's MAE shows that, on average, it is 24.2 hours off the correct number of engineering hours: an improvement of 1.1 hours compared to the current manual quotation process at EeStairs. The MSE score showed a similarly slight improvement. However, the MAPE score shows that our model's predictions are on average 54.7% off from the actual number of engineering hours, whereas the old method is only 34.8% off on average.

When inspecting the performance of our model on a project-by-project basis, we found the largest percentage errors all belonged to projects with less than 30 engineering hours (as seen in Figure 24). The reason for this is likely that our model was optimised for MSE, thus favouring the accuracy of large projects. This is because errors of small projects, while being rather small in absolute terms, could still be large relative to its actual score.

Additionally, we observe that our final Lasso model only uses one feature: the number of balustrade segments. Staircase features are completely excluded by the final model, while the engineering time for balustrades is far lower than that for staircases.. Though this could partially be due to the L1 regularisation applied in Lasso (shrinking some coefficients to 0), models without regularisation or with L2-regularisation generally also used only 1 or 2 features.

The above indicates that the engineered features, as described in Section 3.1.4, have a low predictive power. That Lasso outperformed the other models likely has to do with its tendency towards simpler, more reliable models. This is reflected by its relatively small 6.77% MSE test score increase over the train score, whereas the other models had out-of-sample scores that were between 14.68 and 19.29% higher.

With these observations and interpretations in mind, let us return to our research question:

"How can EeStairs use machine learning to improve its cost estimation process?"

We presented an approach for EeStairs to integrate machine learning into their quotation process. By aggregating data from several sources, followed up with heavy data engineering, we presented a Lasso regression model that showed decent performance across 53 projects at EeStairs. For several reasons, however, it appears this model is not good enough for EeStairs to use in order to improve its cost estimation process.

The first and foremost reason for this is that its performance is only decent compared to EeStairs' manual estimations, while one of the main problems associated with this old method was that the company deemed its performance to be unsatisfactory. Any new method should thus, at the very least, perform significantly better than the old method. Another drawback to this model is the fact that it only uses one feature, besides other limitations which we will discuss in Section 5.2.

The results certainly confirm that cost estimation for Versatile Manufacturing Companies is a challenge, as several researchers have previously noted (e.g. Kingsman et al., 1996; Bashir & Thomson, 2001). Through building our machine learning approach, we stumbled upon several challenges related to infrastructure and data collection, and acquired valuable insights on how to work towards a machine learning pipeline at EeStairs that is accurate, fast and reliable. These insights will be discussed in the recommendations section.

As we have noted previously, most research into cost estimation for manufacturers of custom-made products is conducted at companies that either produce at a large scale, produce products with limited variability, or both. To researchers working at companies with a versatility and scale similar to EeStairs, the recommendations outlined in Section 5.3 should prove valuable in order to identify and overcome obstacles to improving their cost estimation process.

5.2 Limitations

The present research is subject to several noteworthy limitations, encompassing issues of data quantity, quality, and reliability, as well as concerns regarding the application of machine learning techniques and the defined scope of the study.

First, the number of projects included in this research is only 53, due in part to the time-intensive process of data collection. Another reason for this low number is that, while costs are currently estimated both by type of work (e.g. engineering, production, installation) and by project part (per staircase or balustrade), the actual costs made or hours worked is an aggregate either for the entire project, or only separated by the type of work (i.e. it is known which employee worked on which project, but not on which part of it). As such, extensive feature engineering (detailed in Section 3.1.4) was required in order to arrive at aggregate features per project. Significant information loss is likely to have occurred as a result, both in quantity and quality, and the features related to the staircase were the hardest to engineer properly. This is confirmed by the final model, showing that the predictive power of the engineered staircase-related features is low.

Second, to compensate for the low amount of data, the choice was made to include all 53 projects for training and testing (using k-Fold Cross Validation), thus not utilising any unseen data to validate the final model. For similar reasons, the choice was made to not exclude the two potential outlier projects observed in Figure 13, because there is a clear need for the company to be able to predict large projects well. Both could have a slight influence on the reliability of the results.

Third, while working on the research, it became clear data reliability is another limitation: Both the original and actual numbers of engineering hours are subject to potential subjectivity. Interviews with employees showed that factors such as the likelihood of a client accepting an offer may significantly impact the quoted number of hours budgeted for a project. Salespeople may adjust the quote to increase the chances of a sale or adjust it upwards if it is deemed cost is not a significant concern for the client.

Similarly, it is likely that the recorded number of hours worked on projects is biased. Interviewees showed engineers could, for example, step up their productivity when approaching the quoted number of hours, or take a more relaxed approach otherwise. Likewise, they could decide to record the number of hours worked differently, for example by rounding their hours up or down based on the above-mentioned factors. For these reasons, it is likely that the recorded number of engineering hours is biased towards the estimated number, thus artificially increasing the accuracy of the company's previous estimation method.

Finally, because of the necessity for a solution to be easily interpretable by EeStairs' employees (and possibly customers), this research only looked at (linear) regression analysis. Other machine- or deep learning methods (such as Neural Networks) might offer better performance, although the data quality is likely to remain a limiting factor.

5.3 Recommendations

From the points made earlier, it becomes clear this model is not a viable replacement for EeStairs' current cost estimation process. Regardless, this research can provide valuable insights on how machine learning can improve the cost estimation process of EeStairs and companies like it. In this section, we outline four recommendations for the company, followed by additional remarks for researchers at other Versatile Manufacturing Companies.

1. Record actual costs and hours in more detail

Recording costs in higher detail, including materials and work hours, would greatly enhance the data quality of EeStairs. To achieve a balance between practicality and data quality could probably be struck if hours and costs per project part (such as per balustrade or staircase), where feasible. This should include engineering hours.

2. Implement and record machine learning-useable project features

Although at this time EeStairs' data lacked some of the quality and quantity needed for machine learning to make significant improvements to the accuracy of their cost estimation, this is likely to change in the future. As the company implements their new quotation system, they are implementing a system that meticulously records the input variables that help determine projects' costs. These input variables, if recorded over several years of work, could form the basis for a machine learning-project that:

- Requires little time to create and transform data.
- Has much more data to work with.
- Uses data that more closely reflects reality.

3. Increase the objectivity and consistency of both input and output data

The aforementioned points are contingent upon the crucial prerequisite that the data utilised are as unbiased as possible, and evaluated consistently across time. To this end, the company should firmly distinguish between the estimated cost and the quoted price (i.e. the costs should be estimated purely on the projects' details, without thinking of customer expectations in terms of cost). To protect the integrity of this objective data, the company should implement a *subjective* factor, e.g. 'profit margin' or 'discount', that can give the customer a lower or higher price for sales-technical reasons. Additionally, EeStairs should encourage employees to report their hours as accurately as possible, and take away any incentives to do otherwise:

- Judge employees by their overall, long-term performance and do not discipline or reprimand them when they go over time on a single project, to ensure they do not see the quoted number of hours as a depository to report any hours to.
- Give employees the flexibility to report some of their hours to non-project-related tasks (such as team meetings), to make sure they do not report unrelated hours to specific projects.
- Distinguish between productive and unproductive/erroneous hours worked on a project, for instance, in case someone made a mistake and their work needs to be redone.
- In general, make (accurately) reporting hours as simple as possible.

4. Until then, use other machine learning or traditional methods

To improve the results with the current data, other machine learning methods could be explored. For instance, Tayefeh Hashemi et al. (2020) found that, other than regression analysis, artificial neural networks (ANN) are often used for cost estimation, while Badawy (2020) found superior results in cost estimation for residential buildings using an approach combining ANN and regression models. Similarly, it is worth exploring the usage of traditional cost-estimation methods such as those mentioned by Niazi et al. (2006) and García-Crespo et al. (2009), or using a combination of machine learning and traditional methods - for instance, using the model of this research to estimate balustrade costs, and using a knowledge-based system for staircases.

For other VMCs it is advisable to scrutinise the quality and quantity of their data, as many of the problems encountered in this thesis are likely to apply to other VMCs as well. Only when sufficient data of adequate detail is available should companies consider applying machine learning to improve their cost estimation systems. If the company in question has good data on part of the cost, but limited or low-quality data on other parts, companies could consider using a mixed-method approach with machine learning and, for instance, heuristics. If not, we would recommend VMCs to use non-machine learning methods, or to first undertake a project aimed at improving the quality of their existing or future data and data infrastructure.

6 Conclusion

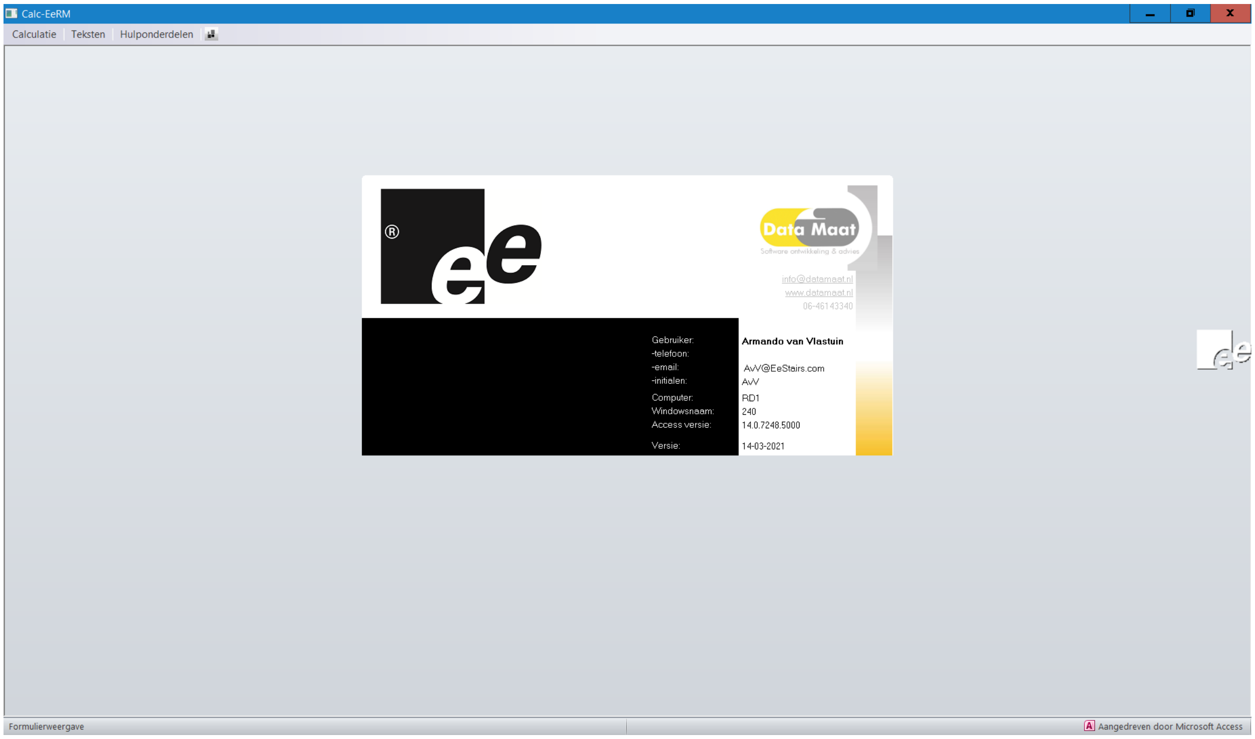
To conclude, the company identified a need to improve and automate their quotation process. As part of this project, we investigated the use of machine learning to predict their projects' engineering costs. Based on interviews and existing literature on cost estimation and design & engineering costs, we selected and created relevant data on 53 projects from EeStairs' existing databases: a relatively small amount of data, due to a time-intensive process of data collection. Because the output data was only available as an aggregate over projects (whereas the input data was broken down by staircase or balustrade), heavy feature engineering was then required.

Aiming to construct a model that was both understandable and implementable by the company, we implemented and optimised four different regression models and selected the best one–the Lasso Regression mode–to proceed with our analysis. While our model's performance was slightly better than EeStairs' manual method based on the mean squared error, it showed a 57% worse accuracy in terms of percent error, mostly because of large errors for projects with fewer than 30 engineering hours. In addition, we found that the optimal model used only one feature, indicating that many of the engineered features offered little predictive value.

From the above results, we concluded our model is not accurate enough to offer a reliable improvement over the old method. We identified several underlying reasons our model is underperforming and offered four core recommendations to EeStairs. Taken together, we took the first steps towards a data-driven approach for estimating costs based on project features at the company. We hope this thesis provides a helpful foundation to further automate and improve EeStairs' cost estimation process, as those of other versatile manufacturing companies.

Appendix A: TrapCalc: EeStairs' current cost estimation program

In this appendix we describe the current cost estimation process, based on unstructured interviews with two of EeStairs' calculators and a demonstration. The current system used to estimate costs and generate a quote is named 'TrapCalc', Dutch and shorthand for 'Staircase Calculator'. The program's start page is shown in Figure 25. By hand of one project whose calculation has been completed, we will illustrate its functioning, censoring sensitive numbers and customer/employee identifiers.



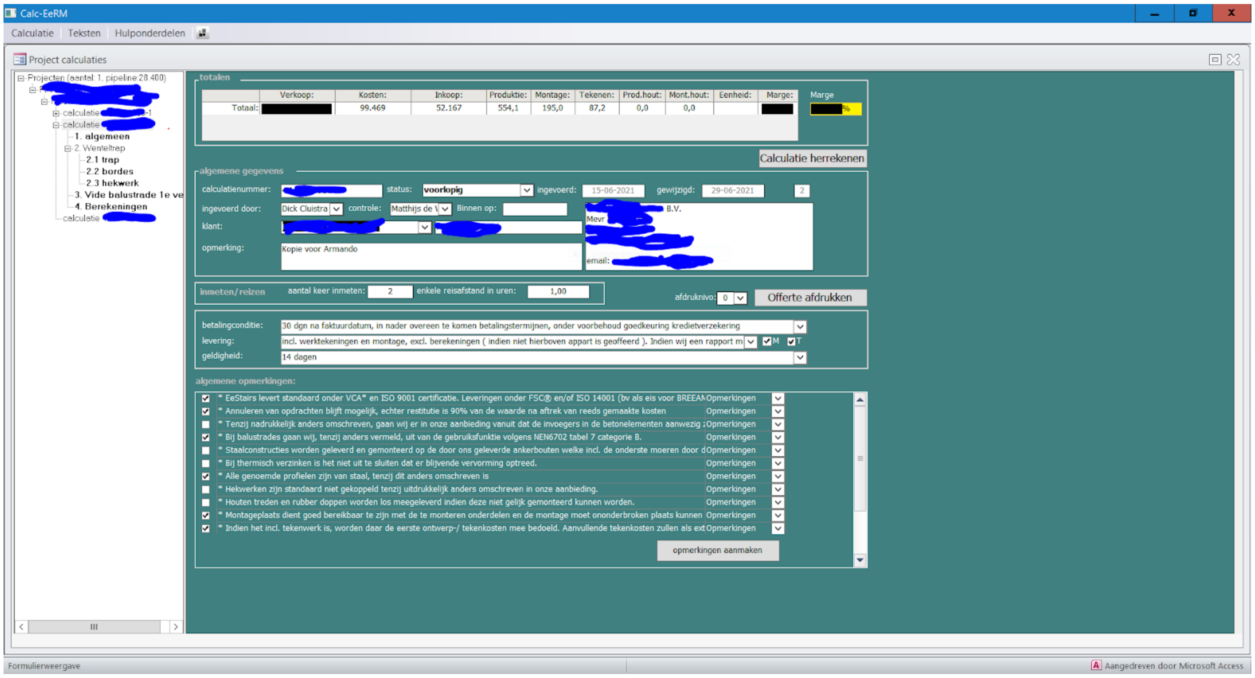
info@datamaat.nl
www.datamaat.nl
06-46143340

Gebruiker: Armando van Vlastuin
-telefoon:
-email: AvV@EeStairs.com
-initialen: AvV
Computer: RD1
Windowsnaam: 240
Access versie: 14.0.7248.5000
Versie: 14-03-2021

Figure 25: Screenshot of the TrapCalc start page.

Overview

After importing some basic project information from the CRM system (such as the customer's name and location), a quote can be created. The first screen to open is the quote overview. After the estimations for all parts of a project have been completed, this overview looks as seen in Figure 26. In the top section, 'totalen', we find the total estimated cost for several categories (including 'tekenen', or engineering), and the quoted price. In the section below is general information about the calculator, the quote, the project and the customer. Next is information on travel distance & the number of times a measurement must be made on location. Then the payment terms, delivery details and the quote's expiration date is noted, and finally a list of general comments (such as terms & conditions) is generated based on which boxes the calculator ticks.



Projecten (aantal: 1, pipeline:28.400)
[redacted]
calculatie [redacted]-1
calculatie [redacted]
1. algemeen
2. Wenteltrap
2.1 trap
2.2 bordes
2.3 hekwerk
3. Vide balustrade 1e ve
4. Berekeningen
calculatie [redacted]

totalen

	Verkoop:	Kosten:	Inkoop:	Produktie:	Montage:	Tekenen:	Prod.hout:	Mont.hout:	Eenheid:	Marge:
Totaal:	[redacted]	99.469	52.167	554,1	195,0	87,2	0,0	0,0		[redacted]

Marge [redacted]%

algemene gegevens

calculatienummer: [redacted] status: voorlopig ingevoerd: 15-06-2021 gewijzigd: 29-06-2021 2
ingevoerd door: Dick Cluistra controle: Matthijs de \ Binnen op:
klant: [redacted] [redacted]
opmerking: Kopie voor Armando

[redacted] B.V.
Mevr [redacted]
[redacted]
email: [redacted]

inmeten/reizen aantal keer inmeten: 2 enkele reisafstand in uren: 1,00
afdruknivo: 0

betalingconditie: 30 dgn na faktuurdatum, in nader overeen te komen betalingstermijnen, onder voorbehoud goedkeuring kredietverzekering
levering: incl. werktekeningen en montage, excl. berekeningen (indien niet hierboven appart is geoffeerd). Indien wij een rapport m ☑M ☑T
geldigheid: 14 dagen

algemene opmerkingen:

☑ * EeStairs levert standaard onder VCA* en ISO 9001 certificatie. Leveringen onder FSC® en/of ISO 14001 (bv als eis voor BREEAN Opmerkingen
☑ * Annuleren van opdrachten blijft mogelijk, echter restitutie is 90% van de waarde na aftrek van reeds gemaakte kosten Opmerkingen
☐ * Tenzij nadrukkelijk anders omschreven, gaan wij er in onze aanbieding vanuit dat de invoegers in de betonelementen aanwezig z Opmerkingen
☑ * Bij balustrades gaan wij, tenzij anders vermeld, uit van de gebruiksfunktie volgens NEN6702 tabel 7 categorie B. Opmerkingen
☐ * Staalconstructies worden geleverd en gemonteerd op de door ons geleverde ankerbouten welke incl. de onderste moeren door d Opmerkingen
☐ * Bij thermisch verzinken is het niet uit te sluiten dat er blijvende vervorming optreed. Opmerkingen
☑ * Alle genoemde profielen zijn van staal, tenzij dit anders omschreven is Opmerkingen
☐ * Hekwerken zijn standaard niet gekoppeld tenzij uitdrukkelijk anders omschreven in onze aanbieding. Opmerkingen
☐ * Houten treden en rubber doppen worden los meegeleverd indien deze niet gelijk gemonteerd kunnen worden. Opmerkingen
☑ * Montageplaats dient goed bereikbaar te zijn met de te monteren onderdelen en de montage moet ononderbroken plaats kunnen Opmerkingen
☑ * Indien het incl. tekenwerk is, worden daar de eerste ontwerp-/ tekenkosten mee bedoeld. Aanvullende tekenkosten zullen als ext Opmerkingen

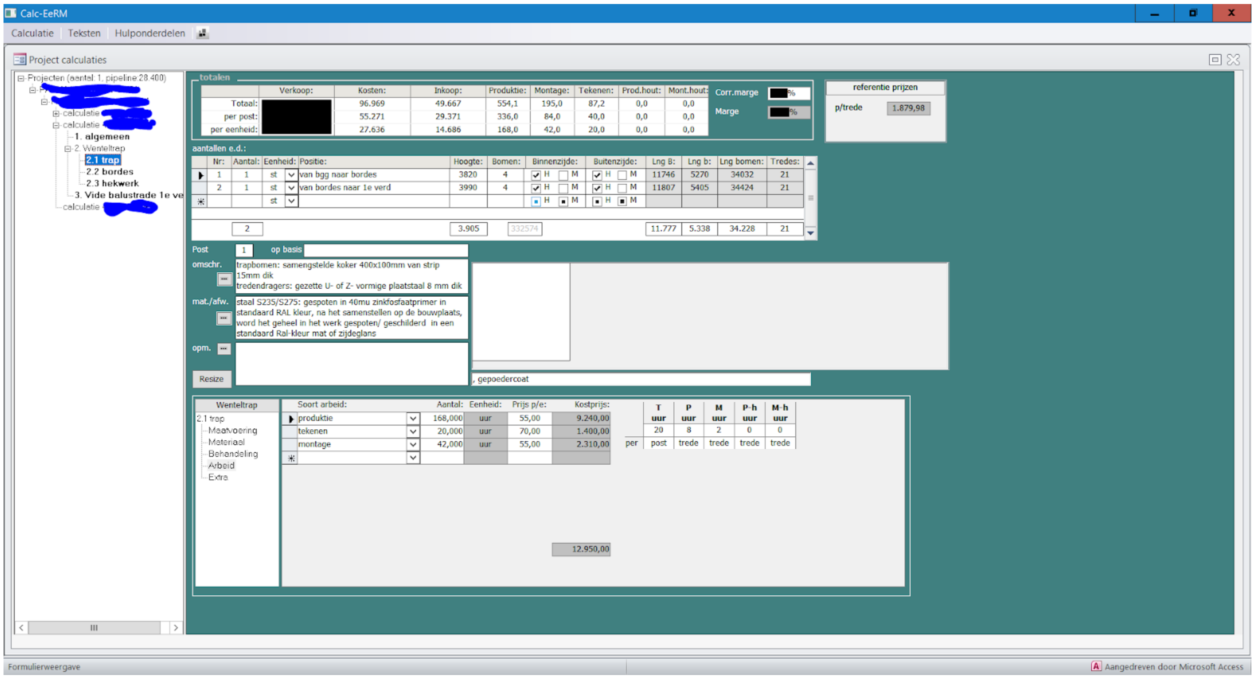
Figure 26: Screenshot of the quote overview.

Estimation per product part

To start calculating the cost for a project, a submenu is created for each part. As seen on the left side of Figure 26, in this project this is divided into ‘algemeen’ (general), ‘wenteltrap’ (spiral staircase), ‘balustrade’ and ‘berekeningen’ (structural calculations). General includes the time spent on taking measurements, installing the staircase, and on transportation, while the structural calculations simply refers to a predetermined price depending on how many different (parts of the) staircase(s) have to be calculated.

The spiral staircase is further divided into ‘trap’ (staircase, the steps), ‘bordes’ (landing, the flat area in the middle of a staircase), and ‘hekwerk’ (railing, the two sides of the staircase), making calculations for the time spent on production and installation per step (of a staircase) or per metre (of a railing), for the materials used, the design time, and any finishing treatment. These calculations are shown in Figure 27. At the bottom is where the calculations for the size, materials, treatment, labour & extras are added. Of these, the size, materials and treatment can be estimated quite accurately: the approximate measurements, types of material to be used and any treatment required can be predicted based on the first designs. After the number of steps is calculated, the approximate materials used per step and for the beams (holding the steps together) are added. Treatment is added at a known price per square metre, since it is normally outsourced.

For the balustrade, in this case the calculation is rather easy, as is often the case: because EeStairs regularly produces similar balustrades, a template is used to which only minor edits are made (such as its length or a specific finish). If this project had a custom balustrade, some more calculations would have to be made in terms of its type of material and the amount of material per metre of length, as well as estimations on the engineering time.



totalen

	Verkoop:	Kosten:	Inkoop:	Produktie:	Montage:	Tekenen:	Prod.hout:	Mont.hout:
Totaal:	[redacted]	96.969	49.667	554,1	195,0	87,2	0,0	0,0
per post:	[redacted]	55.271	29.371	336,0	84,0	40,0	0,0	0,0
per eenheid:	[redacted]	27.636	14.686	168,0	42,0	20,0	0,0	0,0

Corr.marge [redacted]%

Marge [redacted]%

referentie prijzen

p/trede 1.879,98

aantallen e.d.:

Nr:	Aantal:	Eenheid:	Positie:	Hoogte:	Bomen:	Binnenzijde:	Buitenzijde:	Lng B:	Lng b:	Lng bomen:	Tredes:
1	1	st	van bgg naar bordes	3820	4	☑ H □ M	☑ H □ M	11746	5270	34032	21
2	1	st	van bordes naar 1e verd	3990	4	☑ H □ M	☑ H □ M	11807	5405	34424	21
		st				■ H ■ M	■ H ■ M				
	2			3.905	332574			11.777	5.338	34.228	21

Post 1 op basis

omschr. trapbomen: samengstelde koker 400x100mm van strip 15mm dik
tredendragers: gezette U- of Z- vormige plaatstaal 8 mm dik

mat./afw. staal S235/S275: gespoten in 40mu zinkfosfaatprimer in standaard RAL kleur, na het samenstellen op de bouwplaats, word het geheel in het werk gespoten/ geschilderd in een standaard Ral-kleur mat of zijdeglans

opm.

Resize

, gepoedercoat

Wenteltrap

2.1 trap
- Maatvoering
- Materiaal
- Behandeling
- Arbeid
- Extra

Soort arbeid:	Aantal:	Eenheid:	Prijs p/e:	Kostprijs:
produktie	168,000	uur	55,00	9.240,00
tekenen	20,000	uur	70,00	1.400,00
montage	42,000	uur	55,00	2.310,00
				12.950,00

	T uur	P uur	M uur	P-h uur	M-h uur
	20	8	2	0	0
per	post	trede	trede	trede	trede

Figure 27: Screenshot of the calculations done for the staircase.

Estimating engineering time

For brevity's sake, we will not go into detail of processes such as the production & installation, and focus on engineering ('tekenen' in Figure 27). Within this program, for the steps and for the landing, an estimate is made for the amount of time taken to design the entire thing. For balustrades and railings this design time is accounted for per metre. TrapCalc then adds up all these hours spent, as shown in the top table of the above picture.

Degree of automation

Though TrapCalc is not entirely a manual system, the built-in functions to speed up the process are very basic. Aside from simple multiplication or addition of inputs, the main speed-increasing function is the ability to create templates for common products (such as balustrades, some of which EeStairs produces regularly) and terms & conditions, decreasing the time spent typing or calculating in those instances. Additionally, customer data is automatically copied to TrapCalc from the CRM system, and after filling in all costs a quote can be produced with 1 click, including layout & information about EeStairs.

Appendix B: Systematic Literature Review protocol

Search strategy

The question answered through this literature review is the following:

> What are cost drivers for design and engineering?

Databases

For this literature review, the database used is Scopus. This is used both because of its quality as a very extensive multidisciplinary database, and because the University of Twente provides access to it.

Search terms and strategy

To answer this question, a number of key concepts and their synonyms were defined. These can be seen in Table 9.

Key concepts	Synonyms	Narrower	Broader
Cost		Time, effort	Price
Estimation	Estimating	Drivers, estimators, factors	Quotation
Design	x	Drawing, architecture	Engineering

Table 9: Search terms.

The first searches using these terms resulted in thousands of sources, and as such the search query was refined until a large percentage of the remaining sources seemed relevant. This resulted in a query that combined “design” + “cost estimation” and “estimating” + “design cost” and their (most relevant) synonyms. The result of this can be seen in Table 10.

Search query	Database	Hits
((Design OR engineering) AND (“time estimation” OR “estimating time” OR “estimating effort” OR “estimating price” OR “cost drivers” OR “price quotation”)) OR ((Estimating OR estimation) AND (“design time” OR “design effort” or “design cost” OR “engineering time” OR “engineering effort” or “engineering cost”))	Scopus	121

Table 10: Search query.

Inclusion criteria

- Sources that name or identify general cost drivers (whether in terms of time, effort or price) for design and/or engineering

Exclusion criteria

- *Paid sources*, except those accessible via the University of Twente or through Sci-Hub. For budgetary reasons, and because most sources can be found that way.
- Sources that identify cost drivers for a specific unrelated products' design or engineering, such as software engineering or die casting

Judging the articles found on these criteria resulted in the following selection process, seen in Table 11. After a preselection judging whether or not the article is likely to contain information on cost drivers, 26 articles were chosen for reading. Of those that could be accessed some were removed because no factors could be found, others because the factors were very specific to a certain product and as such not relevant for this review. A number of the remaining articles were by the same authors, leading to the removal of the least relevant overlapping articles (often case studies). In one case, the authors' most relevant article on the topic was not included in the literature review but was used as a source in each of their other articles, leading to its inclusion in this literature review. In another case, the factors used came directly from another article, which was subsequently included.

Total number of hits	**121**
Removing duplicates	-3
Selecting based on title	-64
Selecting based on abstract	-28
Article not accessible	-7
Removed after reading	-17
Added after reading	2
Total selected of review	4

Table 11: Resulting articles after several selection steps.

These four articles are introduced and discussed in Section 2.2.

Appendix C: Feature selection process

Synthesis of the systematic literature review

In this section, we select cost drivers from our systematic literature review (SLR) in Section 2.2 that are informative for our use-case and goal. The reasons to exclude a cost driver can be found in Table 12.

#	Criterion	Reasoning
E1	Universality	This factor is the same for all of EeStairs' projects, and as such offers no useful information for our data analysis.
E2	Scope	Assessing the value of this factor is too complex or time-consuming for this project, or the data is simply not available.
E3	Inconsequential	In interviews with Engineers, this factor was found to likely be of little to no influence to the process.
E4	Intransferable	The cost influence of this factor can or should not be transferred to the client.
E5	Redundant	This cost category is already (sufficiently) covered by other factors.
E6	Inapplicable	This cost factor does not apply to EeStairs' projects.

Table 12: Exclusion criteria.

In Table 13, these exclusion criteria are applied to the cost drivers identified in the SLR, and you can find which cost drivers we will include in further research.

Category	Cost driver	Exclusion
Product characteristics		
Product complexity	Structure	Included
	Size	
	Shape	
	Degree of change	
Added demands	Technical difficulty	
	Severity of requirements	

	Use of new technology	E5 - Insofar there are new technologies, these are included in ‘Technical difficulty’ and ‘Severity of requirements’.
	Number of design interdependencies, technologies and sub-assemblies	E5 - Is included in ‘Technical difficulty’ and ‘Severity of requirements’.
Design process		
Use of a formal process	Standardisation	E1/E4 - Design processes are generally the same across projects, and insofar there is a difference it should not be transferred to clients.
	Process control	
Concurrency	Concurrency	E3/E4 - How many different projects are running at the same time was found not to be consequential. Even if it were, this is not a cost that should be transferred to clients.
Design team		
Collaboration	Management complexity	E1 - EeStairs’ team and their collaboration/communication is the same across all projects.
	Team size	
	Methods of communication	
Other	Experience	E2/E4 - Which specific engineer works on a project does differ, but this is beyond the scope of this project. Additionally, the price of a project should depend on the projects’ complexity, not on whether or not the fastest-working engineer happens to be assigned to the project.
	Skill	
	Dedicated spirit	
Design conditions		
Past information	Available data	Included
	Presence of a reference job	Included
Other	Use of design (assisted) tools	E1/E4 - Generally, the same tools are used across projects. Although different engineers use different CAD applications due to personal preference/skill, this (probably marginal) cost difference should not be transferred to the client.

	Management support	E1 - EeStairs’ management is the same across all projects.
	Number of regulations and standards	E2 - Although regulations across country boundaries do influence the project cost, the decision was made to focus on The Netherlands for this project.

Table 13: Applying the exclusion criteria to the cost drivers identified in Section 2.2.

One important difference between EeStairs’ engineering process and those analysed in the SLR, however, is that EeStairs’ process is highly linked to the client and their needs. As such, while the factors in the ‘Design team’ category are all excluded, we will introduce a new and related category as it applies to EeStairs: ‘Client collaboration’. These cost drivers and their reasons for in- or exclusion can be found in Table 14.

Cost driver	**Included?**	**Reasoning**
Management complexity	No	E6/E5 - This is partially inapplicable to EeStairs, and the relevant parts are mostly covered by the following cost driver.
Team size	Yes	In interviews with the company, the size of the clients’ team was found to be quite influential to the decision-making process. Whereas a single decision-maker is relatively easy to deal with, larger teams take longer to make decisions and may require more information.
Methods of communication	No	E1/E3 - The methods of communication are generally similar across projects, and insofar there is a difference it is not consequential.
Experience	Yes	The experience/skill of the decision-maker(s) is thought to be of importance, because experienced clients are easier to work with and make decisions faster.
Skill		
Dedicated spirit	No	E1/E2 - One could assume that all clients who purchase from EeStairs are dedicated at least a little. However, insofar there is a difference here, it would be very hard to determine a clients’ dedication, and definitely outside this projects’ scope.

Table 14: Additional cost drivers and their reasons for in- or exclusion.

EeStairs’ proposed factors

At the start of this project, EeStairs’ director had created a list of potential factors to be used for all aspects of EeStairs’ projects, including engineering. In this section we will lay those relevant to engineering next to the factors synthesised from the SLR, and briefly discuss their overlap and/or differences. You can find this in Table 15.

SLR factor	EeStairs factors	Discussion
Structure	Design & visualisations; Workshop drawings	Although the SLR factors are more specific, they do impact the time spent on design, visualisation and workshop drawings.
Size		
Shape		
Degree of change		
Technical difficulty	Site survey	Depending on the complexity of the product, and especially its installation (and the surrounding environment), there may be a need for multiple site surveys.
Severity of requirements	Structural calculations	Some clients require structural calculations to be performed.
Clients' team size	Project management; Design meetings	The team size and experience are thought to impact both the time it costs in terms of project management, and the number of design meetings necessary to reach a decision.
Clients' experience		
Available data	Building Information Modelling (BIM)	Some clients model their entire building environment, reducing the need for EeStairs to create that part of the model.
Presence of a reference job		Although EeStairs' 'Design & visualisations' and 'Workshop drawings' are placed above, the presence of a reference job does influence the time it costs to complete both and as such, they are applicable here.

Table 15: Discussion of the overlap between the factors resulting from the SLR and those proposed by EeStairs.

Appendix D: Model development process

For the development of our models, we used an approach (experiment A) that started with the four models referenced in Section 3.2.5 in their default scikit-learn setup, using all features as described in Section 3.1.4. In experiment B, we added a feature selection method as described in Section 3.2.4. And finally, in experiment C, we further optimised the Ridge, Bayesian Ridge and Lasso Regression models in order to arrive at our final models, the scores of which are shown in Section 4.1.

All models are trained and tested with a 10-fold Cross Validation procedure, as described in Section 3.2.2. The entire procedure is instantiated 10 times for each model, using 10 different seeds for the CV procedure, in order to increase the comparability between models (due to the small amount of data, it appeared to make quite a large difference depending on how the 10-fold split was seeded). After scaling, optional feature selection and training, the models are run on the test set and compared based on their average MSE, MAE and MAPE scores across splits and seeds. In addition, two scores are used to estimate the variance of our model: the average percentage increase from the train MSE to the test MSE, and the variance of test MSE scores across the 10 CV seeds. Highlighted in **bold** are the best-performing scores in a column.

Experiment A: Basic models

Training our models without any feature selection and with their default setup shows that Lasso performs best, as seen in Table 16. This is in line with the feature eliminating behaviour described in Section 3.2.5.

Model	**MSE**	**MAE**	**MAPE**	**MSE % increase over train score**	**MSE variance**
LinearRegression	1131.37	27.180	0.5866	51.65	2139.97
Ridge	1122.45	27.123	0.5891	50.42	1741.20
BayesianRidge	1135.14	27.217	0.6135	43.34	1575.42
Lasso	**1024.12**	**25.643**	**0.5592**	**34.90**	**1255.45**

Table 16: Error scores and variance of basic models.

Experiment B: Basic models + feature selection

Still using the basic models, in experiment B we then include a SelectFromModel feature selection method, which also uses the default model setup. The results in Table 17 confirm what was seen in experiment A: all models show an improvement due to the feature selection method, except for Lasso - which gives roughly equal results.

Model	MSE	MAE	MAPE	MSE % increase over train score	MSE variance
LinearRegression	970.51	24.996	0.5350	19.29	1281.85
Ridge	959.73	24.836	0.5346	20.78	1472.24
BayesianRidge	**928.18**	**24.450**	**0.5265**	**16.03**	**763.66**
Lasso	1024.11	25.643	0.5592	34.90	1255.51

Table 17: Error scores and variance of basic models, including feature selection.

Experiment C.1: Ridge optimisation

To optimise Ridge Regression, we experimented with different values for the coefficient that multiplies the L2-regularisation score, called Alpha in scikit-learn. Original experiments started with Alpha = [0.1, 1, 5, 10, 20]; 6 and 8 were included after it showed that the score was roughly equal between Alpha = 5 and 10. Ultimately, Alpha = 8 showed the best results, as can be seen in Table 18.

Alpha	MSE	MAE	MAPE	MSE % increase over train score	MSE variance
0.1	966.94	24.882	**0.5323**	21.84	1667.43
1.0	959.73	24.836	0.5346	20.78	1472.24
5.0	945.53	24.746	0.5428	17.17	1261.27
6.0	944.83	24.769	0.5462	16.48	1064.57
8.0	**941.02**	**24.737**	0.5508	14.68	**665.55**
10.0	946.62	24.859	0.5573	13.97	724.00
20.0	1008.44	25.905	0.5932	**13.44**	1174.74

Table 18: Error scores and variance of several Ridge Regression settings.

Experiment C.2: Lasso optimisation

A similar experiment was conducted for Lasso, which showed superior results for alpha = 5. The results can be seen in Table 19.

Alpha	MSE	MAE	MAPE	MSE % increase over train score	MSE variance
0.1	1120.28	27.041	0.5841	50.25	1823.13

1.0	1024.11	25.643	0.5592	34.90	1255.51
5.0	**888.26**	**24.179**	**0.5452**	**6.77**	575.71
10.0	969.24	25.524	0.5860	6.76	**542.15**
20.0	1324.02	29.700	0.6812	9.61	3190.07

Table 19: Error scores and variance of several Lasso Regression settings.

Experiment C.3: BayesianRidge optimisation

Finally, BayesianRidge was optimised. Due to the much larger number of hyperparameters that can be optimised for this method, we opted to use the method BayesSearchCV (scikit-optimize, 2020). BayesSearchCV implements a Bayesian optimisation over the hyperparameters of our BayesianRidge model. The resulting best setup showed no significant improvement over the default settings, as can be seen in Table 20.

Method	**MSE**	**MAE**	**MAPE**	**MSE % increase over train score**	**MSE variance**
Benchmark (no search)	**928.18**	**24.450**	**0.5265**	**16.03**	**763.66**
BayesSearchCV	935.04	24.652	0.5296	16.71	836.45

Table 20: Error scores and variance of several Bayesian Ridge Regression settings.

Experiment C: Final comparison of optimised models

The best-performing Ridge, Lasso and BayesianRidge models were then taken and compared with each other and the default Linear Regression model, as shown in Section 4.1.

Bibliography

Sossi Alaoui, S., Farhaoui, Y. & Aksasse, B. (2018). Classification algorithms in Data Mining. *International Journal of Tomography and Simulation*, *31*, 34-44.

Amaro, G., Hendry, L., & Kingsman, B. (1999). Competitive advantage, customisation and a new taxonomy for non make‐to‐stock companies. *International Journal of Operations & Production Management*, *19*(4), 349-371. https://doi.org/10.1108/01443579910254213

Badawy, M. (2020). A hybrid approach for a cost estimate of residential buildings in Egypt at the early stage. *Asian Journal of Civil Engineering, 21*, 763–774. https://doi.org/10.1007/s42107-020-00237-z

Bashir, H. A., & Thomson, V. (1999). Metrics for design projects: a review. *Design Studies, 20*(3), 263–277. https://doi.org/10.1016/s0142-694x(98)00024-6

Bashir, H.A. & Thomson, V. (2001). Models for estimating design effort and time, *Design Studies, 22*(2), 141-155.

Benedetto, H., Bernardes, M. M. E. S., & Vieira, D. (2018). Proposed framework for estimating effort in design projects. *International Journal of Managing Projects in Business, 11*(2), 257-274. https://doi.org/10.1108/ijmpb-03-2017-0022

Bengio, Y., & Grandvalet, Y. (2004). No Unbiased Estimator of the Variance of K-Fold Cross-Validation. *Journal of Machine Learning Research, 5,* 1089–1105. https://www.jmlr.org/papers/volume5/grandvalet04a/grandvalet04a.pdf

Bernardes, M. M. E. S., Benedetto, H., Chain, M. C., & Vieira, D. R. (2019). Exploring the Context of Price Quotation on Design Projects. *The Journal of Modern Project Management, 7*(1). https://doi.org/10.19255/jmpm01911

Bronstein, A. (2017, May 17). *Train/Test Split and Cross Validation in Python*. Towards Data Science.

https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6

Denkena, B., Lorenzen, L., & Schürmeyer, J. T. (2009). Rule-based quotation costing of pressure die casting moulds. *Production Engineering, 3*(1), 87–94. https://doi.org/10.1007/s11740-008-0139-8

García-Crespo, Á., Ruiz-Mezcua, B., Lopez-Cuadrado, J. L., & González-Carrasco, I. (2009). A review of conventional and knowledge based systems for machining price quotation. *Journal of Intelligent Manufacturing, 22*(6), 823–841. https://doi.org/10.1007/s10845-009-0335-1

Grabenstetter, D. H., & Usher, J. M. (2013). Determining job complexity in an engineer to order environment for due date estimation using a proposed framework. *International Journal of Production Research, 51*(19), 5728–5740. https://doi.org/10.1080/00207543.2013.787169

Gupta, S. (2021, April 26). *RMSE: What does it mean?* Medium. https://medium.com/@mygreatlearning/rmse-what-does-it-mean-2d446c0b1d0e

Heerkens, H., & Van Winden, A. (2021). *Solving Managerial Problems Systematically*. Abingdon, England: Routledge. https://doi.org/10.4324/9781003186038

Hellenbrand, D., Helten, K. & Lindemann, U. (2010). Approach for development cost estimation in early design phases. *11th International Design Conference, DESIGN 2010*, 779-788.

Hvam, L., Malis, M., Hansen, B., & Riis, J. (2004). Reengineering of the quotation process: application of knowledge based systems. *Business Process Management Journal, 10*(2), 200–213. https://doi.org/10.1108/14637150410530262

Kingsman, B. G., Hendry, L., Mercer, A., & De Souza, A. a. U. (1996). Responding to customer enquiries in make-to-order companies Problems and solutions. *International Journal of Production Economics*, *46–47*, 219–231. https://doi.org/10.1016/0925-5273(95)00199-9

Kingsman, B. G., & de Souza, A. A., (1997). A knowledge-based decision support system for cost estimation and pricing decisions in versatile manufacturing companies. *International Journal of Production Economics*, *53*(2), 119-139. https://doi.org/10.1016/S0925-5273(97)00116-3

Koehrsen, W. (2018, April 13). *Introduction to Bayesian Linear Regression*. Towards Data Science. https://towardsdatascience.com/introduction-to-bayesian-linear-regression-e66e60791ea7

Kumar, P. (2008) Design Process Modeling: Towards an Ontology of Engineering Design Activities. *All Theses*, *417*. https://tigerprints.clemson.edu/all_theses/417

Matel, E., Vahdatikhaki, F., Hosseinyalamdary, S., Evers, T., & Voordijk, H. (2019). An artificial neural network approach for cost estimation of engineering services. *The International Journal of Construction Management, 22*(7), 1274–1287. https://doi.org/10.1080/15623599.2019.1692400

Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. Sebastopol, CA, United States: O'Reilly Media, Inc.

Niazi, A., Dai, J. S., Balabani, S., & Seneviratne, L. (2006c). Product Cost Estimation: Technique Classification and Methodology Review. *Journal of Manufacturing Science and Engineering-Transactions of the Asme, 128*(2), 563–575. https://doi.org/10.1115/1.2137750

Peffers, K., Tuunanen, T., Gengler, C. E., Rossi, M., Hui, W., Virtanen, V., & Bragge, J. (2006). The Design Science Research Process: A Model for Producing and Presenting Information Systems Research. *Proceedings of the First International Conference on Design Science Research in Information Systems and Technology,* 83-106. https://doi.org/10.48550/arXiv.2006.02763

Rittel, H. W., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Sciences, 4*(2), 155–169. https://doi.org/10.1007/bf01405730

Salam, A., Bhuiyan, N., Gouw, G., & Raza, S. A. (2009). Estimating design effort for the compressor design department: a case study at Pratt & Whitney Canada. *Design Studies, 30*(3), 303–319. https://doi.org/10.1016/j.destud.2008.10.003

Salam, A., & Bhuiyan, N. (2016). Estimating design effort using parametric models: A case study at Pratt & Whitney Canada. *Concurrent Engineering, 24*(2), 129-138. https://doi.org/10.1177/1063293x16631800

scikit-learn. (2023). *API Reference — scikit-learn 1.2.2 documentation*. Scikit-learn. https://scikit-learn.org/stable/modules/classes.html#module-sklearn.linear_model

scikit-optimize. (2020). *skopt.BayesSearchCV — scikit-optimize 0.8.1 documentation.* Scikit-optimize. https://scikit-optimize.github.io/stable/modules/generated/skopt.BayesSearchCV.html

Sewagu. (2010, November 5). *File:Linear regression.svg*. Wikimedia Commons. https://commons.wikimedia.org/wiki/File:Linear_regression.svg

Tayefeh Hashemi, S., Ebadati, O.M. & Kaur, H. (2020). Cost estimation and prediction in construction projects: a systematic review on machine learning techniques. *SN Applied Sciences, 2,* 1703. https://doi.org/10.1007/s42452-020-03497-1

Xu, D., & Yan, H. (2006). An intelligent estimation method for product design time. *The International Journal of Advanced Manufacturing Technology, 30*(7–8), 601–613. https://doi.org/10.1007/s00170-005-0098-6

Zhang, J., Nault, B. R., Yang, W., & Tu, Y. (2012). Dynamic price quotation in a responsive supply chain for one-of-a-kind production. *International Journal of Production Economics, 139*(1), 275–287. https://doi.org/10.1016/j.ijpe.2012.05.011