1/17/2022

Estimating the white-collar hours with statistical learning

Gijs Verschuur
MASTER GRADUATION PROJECT AT VDL ENERGY SYSTEMS

Author

P.G. (Gijs) Verschuur (s1577417)

VDL Energy Systems

Darwin 10

7609 RL Almelo

(0546) 649 400

Company supervisor

M. (Mark) Smit MSc.

Head operations

University of Twente

Drienerlolaan 5

7522 NB Enschede

(053) 489 9111

First supervisor

Dr. E. Topan

Assistant professor IEBIS

Second supervisor

Dr. I. Seyran Topan

Lecturer IEBIS

Preface

Before you lies my final work as a student. After 7,5 years of studying the end is there. The last year I have spent my time at VDL Energy Systems to jump over the final hurdle; graduating. In hindsight, the process of graduating was not as smooth as I hoped, but I am very happy that unless the pandemic and the hack I am able to present my work. The topic of statistical learning was new for me, but I do believe that the principles behind it will useful later in my life.

I want to thank Bart Bramer, Mark Smit & Casper Heidemann from VES for providing me with the opportunity to graduate at their company and helping me in the process. Next, I want to thank Engin Topan and Ipek Seyran Topan for their guidance from the university point of view. It was not an easy process to come to an end, but the two of you always kept an opportunistic standpoint and provided me with possibilities to finalise my thesis. Someone else who help made the process easier is my girlfriend, so also a big thanks to her. Last, I want to thank my friends Jiska, Thijs & Niek for the fun times throughout the master.

Have fun reading!

Gijs Verschuur

Management summary

This research takes place within the sales and project management department of VDL Energy Systems (VES), Almelo. VES produces compressors and turbines with for example piping and electricity for the oil and gas market. Both departments want to have a better insight into the number of white-collar hours needed per project.

A project is one customer order and can consist of multiple pieces and designs. The insight is necessary because of two reasons. Sales department wants to make more competitive bids in their sales process, while project management cares about the timeline and progress after the handover from sales. A good estimated bid that leads to a sold project decreases the number of budget overruns during operations. Currently, the only way of estimating the white-collar hours, which both departments rely on, is based on expert judgement. For the blue-collar hours, the hours made by workers in the workshop, sales has a tool that estimates the number of hours based on historical data. The goal of this research is to develop such a model of the white-collar hours, to help the sales department in their decision making and ultimately help project management by selling better-estimated projects. This leads to the research question:

How can the quality of the white-hour estimates during the bid phase be increased, focussing on data based on old estimates and finished projects, using statistical learning methods.

The type of projects differs a lot, which makes it hard to use a standard model. Statistical learning (SL) is an option to be able to cope properly with future projects. Within the field of SL, multiple methods provide a way to estimate. However, after the change of holding company in 2018, not all data were set over to the new ERP system accurately. The total number of available projects in the VES ERP system is 45. This is small for an SL study, causing a narrower range of choices of methods. After the literature study, the methods of choice are Multiple Linear Regression (MLR), Decision Tree (DT), and Random Forest (RF). The three methods are coded in the software R and finally formed into a tool which the sales department can use to estimate themselves.

A total of nine departments make up all the white-collar hours. This means that the analysis consists of the building of 27, nine times three, models. Every model has the same inputs, the six feature variables. These are category customer, main component, the sum of all blue-collar hours, pieces, designs, and the intensity level of the required specifications. Every method starts with model selection and to ensure the model does not overfit. The next step is to generate the final 27 models. The root mean squared error (RMSE) is the main evaluation metric to determine how well a model predicts. Further analysis of the DT model consists of determining intervals per endpoint of the tree. Variable importance and quantile regression forest are part of the RF model. These methods show extra insights on the DT and RF models next to the RMSE.

Evaluations of the RMSE show that the prediction power for all departments is shallow, with the engineering department as the hardest to predict. The RMSE also shows that every method is the best performing method at least once. The methods have trouble making well-performing models because of the lack of information. However, comparing the methods result to the estimates VES made, for every department is a model that outperforms the expert judgement of VES. The data driven analysis is a better way to estimate, so the recommendation is to keep adding data to the dataset and to use the models as a conversation starter with the individual departments.

Contents

Table of figures

Table of tables

1. Introduction

Chapter 1 starts with a short company description to get a grasp of the industry and activities. Next, the second section introduces the problem. Followed by a description and goal of the project. The fourth part shows the structure of the report. Finally, the research framework determines the research questions, and the scope gives a quick summary.

1.1 Company description

The description is based on the history of the company, which goes back to two timelines. The timelines are based on the location of the company and of the current owner. After explaining both, the last section shines a light on the existing businesses within VES.

VDL Group

VDL was founded as a family business in 1953. The family van der Leegte, led by Pieter van der Leegte, started in the metalworking business. Over the years, VDL Group acquired multiple companies to broaden its portfolio. Currently, the VDL group consists of 106 firms located in 19 different countries with around 15.000 employees. The current activities of VDL are divided into four divisions, which are supplies, car assembly, busses, and end products. Inhabitants of Twente and Enschede might recognise the company logo, which is on the Syntus busses that provide public transport in this region.

VDL Energy Systems

VDL Energy Systems, or VES, was acquired by the VDL group in 2018. It was the 100th business VDL Group started to invest time and money in. However, the history of the company goes back to 1868. At that time, Stork founded the company. Especially in the city of Hengelo, Stork is a well-known name. Stork was one of the first machine-building companies that heavily influenced the region's economy by providing the local textile industry with the necessary equipment. After an adventure under the banner of Delaval Stork, Mannesmann, Atecs, and Vodafone Airtouch, Siemens took over. They owned the location in the centre of Hengelo from 2003 until 2018. At acquisition, VDL determined the name "Energy Systems" to this location to describe the broad range of activities. In November 2021 the business moved from Hengelo to Almelo.

Current businesses

VES operates in the oil and gas industry. The focus is on the packaging and testing of gas turbines and compressor packages. Other activities include detailed engineering, full load string testing, parts production and service activities. Chapter 2 provides a more in-depth explanation of what those activities are. All those activities are part of the old-fashioned fossil fuel industry. Therefore, VES made the strategic decision to start changing its focus on the energy transition. Part of the strategy is to be the OEM (original equipment manufacturer) of multiple green energy solutions, such as battery containers and fuel cells. Activities regarding the energy transition are upcoming and are currently a small part of the company.

1.2 Problem introduction

This research focuses on the duration of packaging projects within VES. In this research, a project means an order of a customer. Besides packaging projects, VES also has part production projects and new business projects which are not the scope of this research. The recurring problem is that projects of all types take longer than the initial planning and cause budget overruns. Back in the Siemens era, the activities were mainly focused on packaging. This focus implied a project-based way of working with engineering and the possibility of producing necessary parts. When VDL acquired the company, the focus shifted to packaging activities with less engineering. At the same time, the production of parts became a more significant part, forcing a company to take a more product-based approach. Recently, VES again focuses on more engineering heavy projects, but in a new style. However, the implementation of a broadened range of activities is not that easy. VES employees feel that much work is done with the old method that originates from the Siemens era. The main focus point is how projects are sold and handled within the office. Collaboration between departments, the initial planning or timeline of a project, and a missing clear task description were all mentioned as possible shortcomings during the execution of a project. On the other hand, they have the feeling that the production itself is managed well, which leaves the opportunity to look at the number of hours for an entire project in the office.

1.3 Problem description and goal of the project

Multiple factors influence the duration of projects, and therefore if a project is finished in time. There is no one main reason or problem that affects all projects within VES. Per type of project, multiple variables influence the planned project duration and if the deadlines are met. Those variables can include but are not limited to order specifications, number of budgeted hours, involvement of departments, and supplier lead times. Managing those variables is the task of the project managers. It is their responsibility to guard the initial timeline because not delivering on time is costly.

Instead of doing everything in your power to protect the timeline, the question can be raised if the planning was realistic in the first place. When it is not, the problem changes to a planning problem. However, this contradicts with the sales department. They mentioned that quotations or tenders get lost because of the amount of time it costs VES to complete a project. This loss of sales happens for packaging projects, but the same is true for some part production projects. It is a small problem because, with the acquisition of VES, Siemens contractually agreed on keeping to provide the demand. However, when that agreement ends, VES finds itself in a disadvantageous position with hard to sell parts or packaging projects.

The main focus is time, explicitly working hours, which is a type of cost for VES. Zooming in on the costs of packaging projects and some of the part production projects, VES classifies three main sorts of cost: material costs, direct or blue-collar hours, and indirect or white-collar hours. When VES participates in a tender or quotation, these three costs are estimated to form a bid. During the bid phase of a project, it occurs that not all information from the customer is known. However, estimations still have to be made to come up with a bid. Right now, most of the cost estimation for white-collar hours is based on the experience of a department or department leader. However, this is not ideal. Shepperd & Cartwright (2001) already observed different phenomena that could happen when companies estimate through expert judgement. These are:

- A preference for singular as opposed to distributional information,
- Recall impacted by recency and "vividness",
- Distortion of probabilities,
- Anchoring and adjustment,
- Group dynamics and a fear of voicing "negative" opinions.

Because of the disadvantageous position, the sales department started to work on a tool that should provide more insight into how the costs for a specific project are determined. In this tool, they want to indicate where all the costs come from before coming up with a bid to the customer. This tool primarily focuses on checking boxes in specific specifications that indicate material costs and blue-collar hours. However, estimating the number of white-collar hours needed per department is more complex than for blue-collar hours. For example, it cannot be determined by the number of welds that have to be placed or the meters of piping that has to be connected. Next to the difficulty of predicting the white-collar hours, expert judgement could be a problem. Within the sales department, the feeling exists that when a project exceeded the budget, the responsible department leader will increase the estimation of the white-collar hours for the next project. A different possible problem is that the tools used by the individual departments are outdated. For example, if the department engineering determined certain steps to do all the work, the amount of work per step is a fixed number of hours that rarely changes on performance.

Right now, the sales department does not always know if the white-collar hour estimations made by expert judgement are realistic. This also influences the execution after the project is sold. If the sold number of hours badly represents the operations, delays or budget overruns per department become more likely. Therefore, both the sales department and the project management want a more in-depth analysis of how a project is sold.

Problem cluster

A problem cluster helps to provide an overview of how the problems relate to each other. Figure 1 shows the problem cluster of this research.

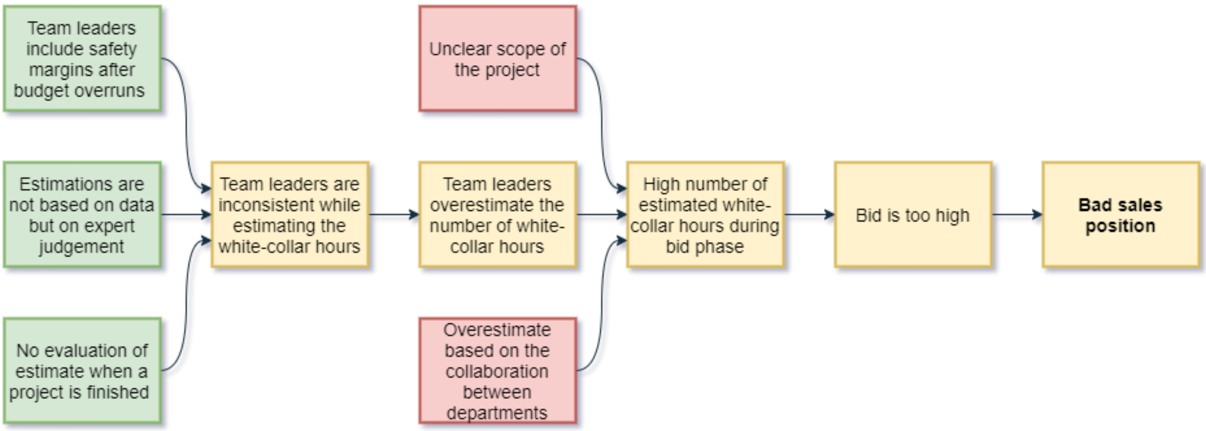

Team leaders include safety margins after budget overruns
Estimations are not based on data but on expert judgement
No evaluation of estimate when a project is finished
Team leaders are inconsistent while estimating the white-collar hours
Unclear scope of the project
Team leaders overestimate the number of white-collar hours
Overestimate based on the collaboration between departments
High number of estimated white-collar hours during bid phase
Bid is too high
Bad sales position

Figure 1: The problem cluster.

Based on the book Geen probleem (Heerkens & van Winden, 2012), the focus should be on the end problems. Figure 1 contains five of them. The difference between colours indicates if the problem is included in this research. Green means included, and red means excluded. The previous section mentioned the first and second green problems. The last problem, no evaluation of estimate when a project is finished, somewhat relates to the safety margins. Because there is no evaluation after the end of a project, a safety margin for the next similar project seems logical from a department perspective. The reason why it is a standalone problem is the way of estimating. The finished projects should always feed new decisions or estimates.

To solve this discrepancy on what happens within the company and what the sales department wants to happen within the company, lesser hours needed for a project, the focus of this research could be on historical data to judge how after the estimation the company as a whole carried out the project. Unfortunately, after the acquisition by VDL, they immediately implemented a new enterprise resource planning (ERP) system. The old servers and hard drives are inaccessible, which makes a data-driven project hard to do. However, the new ERP system slowly gets filled, which means data becomes more and more available. Accordingly, a data-driven approach to estimate the white-collar hours might be possible when more projects are finished.

The problem cluster leads to the problem statement of this research:

The white-collar hours are likely over- or under estimated throughout the bid phase, likely caused by inconsistent expert judgement.

This leads to the main research question.

How can the quality of the white-hour estimates during the bid phase be increased, focussing on data based on old estimates and finished projects, using statistical learning methods.

Therefore the goal is to:

Analyse the current white-collar estimation methods and provide a data-driven model that increases the quality of the bid, based on estimates and analysis.

It is crucial to keep in mind that the focus is on packaging projects. The main reason for this focus is the department from which this study is initiated, which is project management. They are after the estimation, independently of the quality of the estimate, responsible for staying within the budget. A second note is that the research is explicitly focused on hours and not on costs.

1.4 Methodology

This research deviates a little from the standard form. Figure 2 shows the structure.

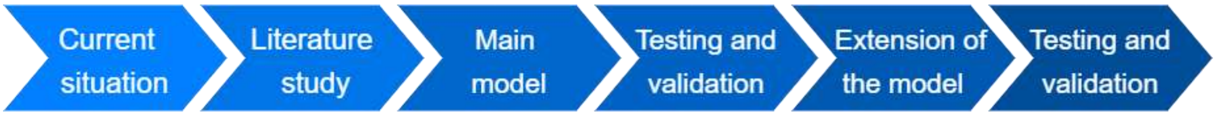


Figure 2: Structure of the research.

The structure differs because of the double modelling. Usually, a model is made and then tested, but a dual modelling approach is chosen because of the lack of data. The first model will focus on estimations from the past. Those are used to form a model which indicates the white-collar hours. Next, the focus of the extension is to learn from finished projects. Although the total number of finished projects is low, some learning from how projects are handled should increase the quality of the model. The deliverables cause this structure agreed on. The models cannot be made in parallel because the main model serves as input for the extension.

1.5 Research framework

Research questions help to reach the goal. The questions are divided per part of the methodology. Since a double modelling approach is chosen, some elements between the models are similar. The literature review will therefore consider both cases at the same time because it is likely that some core theory overlaps.

Current situation

1. How do the departments currently determine how many hours are necessary?
2. What are the factors or cost drivers that influence the number of estimated hours per department?

Main model

3. What method is the most suitable regarding small amounts of data?
4. What are the most suitable methods to estimate the white-collar hours?

Testing and validation of the main model

5. How should the models be validated and tested?
6. What are the best methods to validate and test the main model?

Conclusion

7. What model performs best per department?

1.6 Scope

The scope is to use data from former projects to the quality of the white-collar hours' estimation throughout the bid phase of a project. The lack of data makes it hard to judge the performance, but it should be appropriately handled when more and more data is collected. The scope for the first phase is to evaluate the consistency of the estimations and provide a model that will help sales start a discussion when the estimated number of hours is out of line with earlier projects. The second phase focuses on judging the performance. Therefore a small data model or method is designed to help VES in the future to evaluate if the initially estimated number of hours should be adjusted based on the historical performance during the packaging projects. The model should give a guideline in either a point estimate or a range estimate of how many hours a specific department is expected to spend. The model serves as a tool made specifically for VDL Energy Systems. The main model will be statistical learning. It suits the situation the best and is a broad field where many opportunities lie.

2. Situation

The description of the present situation consists of six parts. Section 2.1 explains the core business of VES to indicate the type of work they do. Section 2.2 explains the cost involved in a packaging project. Next, Section 2.3 focuses on the problematic white-collars for packaging projects. Section 2.4 describes the sales process since the number of white-collar hours is determined in this process. Section 2.5 explains how department estimate right now, and the last section shows the two different types of data that serve as input in the model of Chapter 4.

2.1 Types of projects

Within VES, there are three main categories of project orders. Those are packaging, part production, and new business. The categories are addressed individually.

Packaging

This category involves the historically speaking main business of VES. Packaging within VES means building the product into one piece, for example, an industrial compressor. The main activities of a packaging project include piping, isolation, electrics and assembly. For most of the projects, multiple components are bought or provided by the customers which are installed on a base frame. Figure 3 shows such a machine. The numbers mention different components. The technical information about the numbers is not essential. Instead, they serve as an indication that VES gets provided with, in this case, four main components and builds them together as one piece.

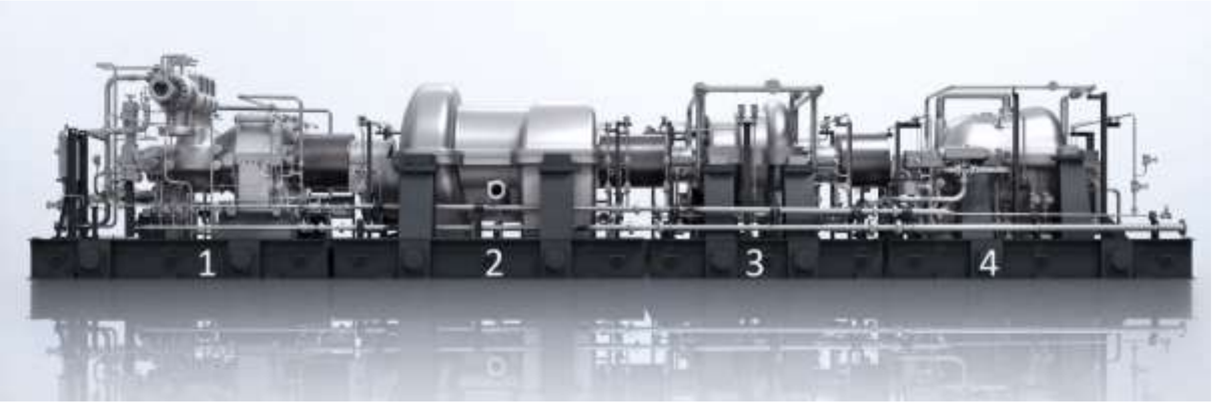


Figure 3: Example of a finished packaging project.

It is hard to say what a standard packaging project looks like because of all the variations in parts, modules, and tasks the customer wants VES to do. However, there are some ways to standardise. For example, the size of the compressor does give some sort of an indication. VES have experience working with a compressor called the SGT. The different types include the SGT-100 up to the SGT-700, where a larger number implies higher power. The term packaging can be inferred from the fact that the entire machine gets packed with pipelines and electric cables. In the past, VES made compressors, which is component 2 in the picture. VES could still make them at this moment in time, but it is rarely part of the scope agreed on with the customer for any packaging project. In other words, that service or type of works does not get sold anymore by VES.

Part production

VES has a workshop for parts since some packaging projects need custom piping and tailor-made parts. Besides making parts for the packaging projects, VES also sells parts to customers. In the workshop, a lot of metalworking can be done. For example, sawing, machining, drilling, bending, and welding. The latter is not part of metalworking but a big part of the company. Examples of parts are axes, fans, labyrinths, diaphragms, volutes, sleeves, thrust collars, balance drums.

New business

With the thought that the oil and gas industry is not forever, VES started with projects to do their part in the energy transition. CO_2 neutral energy solutions are developed or looked into. Examples of such projects are fuel cells or battery containers. Important to note is that VES is the OEM (original equipment manufacturer) in the case of a sold product. This ownership is the same for parts, although the implications are different. Unfortunately, the current state of the majority of projects is still R&D.

2.2 Costing of a packaging project

Three types of costs determine the total costs of a packaging project. Those costs are material costs, blue-collar hours and white-collar hours. To determine the total cost of a project, the blue-collar hours and white-collar hours are multiplied with their rates, respectively. The financial department of VES defined those rates per cost group. A cost group describes a specific role within a department. Section 2.4.1 explains cost groups in detail. Cost group rates differ between per department and even per machine that has to be operated. These rates are excluded in this research since some rates make up for more than only man-hours.

Material costs

Material costs are straightforward. When a particular metal, part or another material is required for a packaging project, it needs to be bought somewhere. VES has contact with over 200 suppliers. It can even happen that the customer provides a list of suppliers because they want to use their materials or components. A specific type of materials within VES are free-issue materials. Free-issue means provided by the customers, which means zero costs for VES. Commonly, a customer ships individual pieces from Figure 3 to Hengelo to let VES build a machine out of it. Free-issue materials are not limited to main components. Free-issue means that VES does not have to invest money in a product or is responsible for the quality.

Blue-collar hours

Blue-collar hours are the hours involved with the metal handling, part production or assembly. The term is based on the standard blue suits you can find the workers in. Another term for blue-collar hours is direct hours. The blue-collar workers add immediate value to a certain product or project, which makes them the company's core. Without those workers, no value can be added.

White-collar hours

White-collar hours, or indirect hours, are hours made by the supportive departments. Examples of those departments are sales, project management, procurement, planning and quality. However, quality is a special case. Therefore, when an ISO norm certification is required, one could argue that the Q&A department provides value and should be classified as blue-collar.

2.3 White-collar hours in packaging projects

The best way to illustrate the involvement of white-collar hours throughout a packaging project is with VES' main process. It is an overview, visualised for white-collar hours in Figure 4, made three years ago that shows the company's flow of a packaging project. Currently, the correctness of the overview is being discussed. So, while it might not be the best way of working, it still shows how the company thought at some point how the involvement per department should be and at what moment in the project.

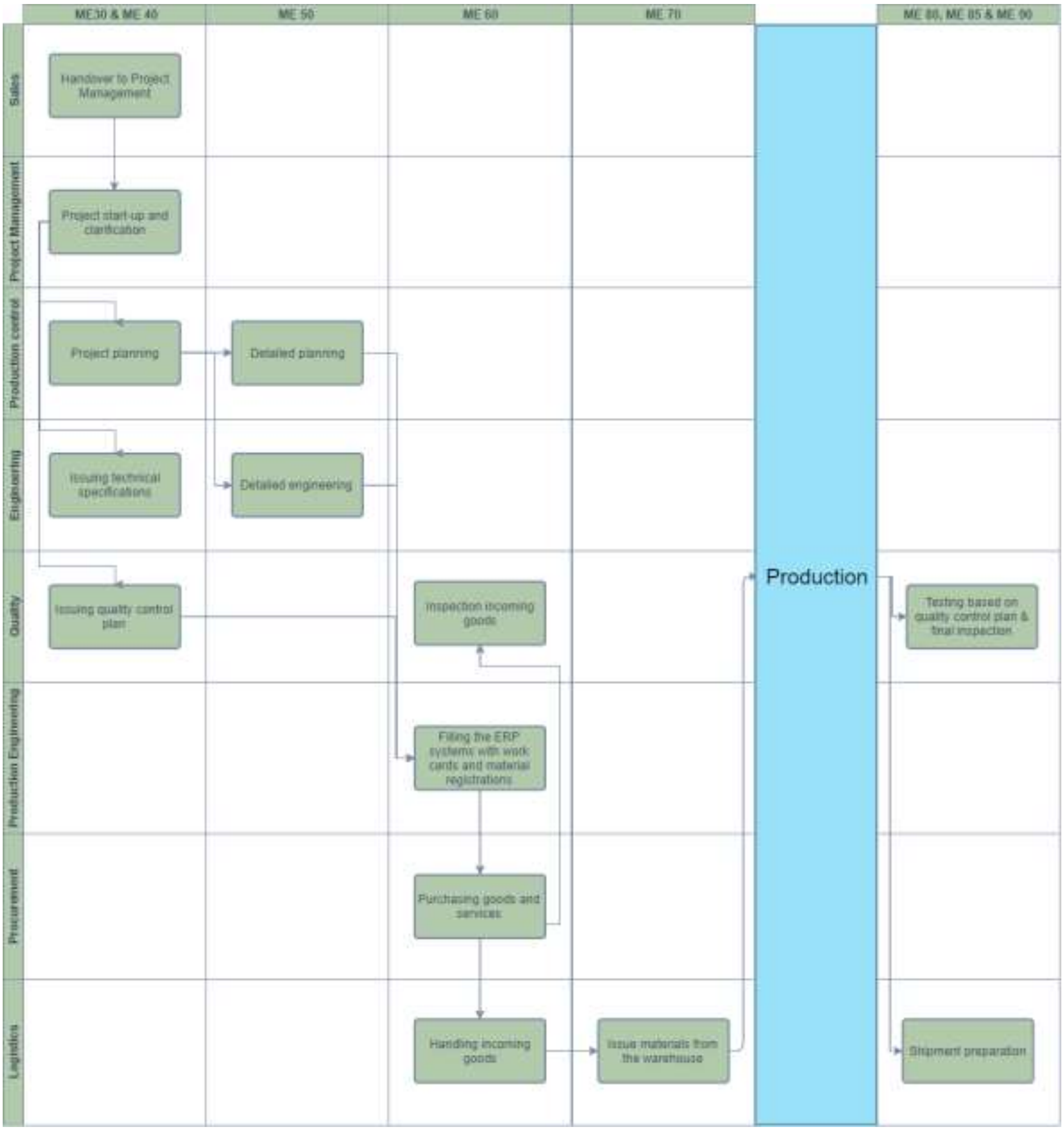

ME30 & ME 40
ME 50
ME 60
ME 70
ME 80, ME 85 & ME 90
Sales
Project Management
Production control
Engineering
Quality
Production Engineering
Procurement
Logistics
Handover to Project Management
Project start-up and clarification
Project planning
Detailed planning
Issuing technical specifications
Detailed engineering
Issuing quality control plan
Inspection incoming goods
Filling the ERP systems with work cards and material registrations
Purchasing goods and services
Handling incoming goods
Issue materials from the warehouse
Production
Testing based on quality control plan & final inspection
Shipment preparation

Figure 4: Main workflow white-collar hours, indicated by green cells.

Figure 4 shows almost all white-collar departments that do their part throughout a project. The vertical lanes are milestones specified by VES. Appendix A shows the steps per milestone. The sales department is only mentioned with the handover. The main reason the sales department is left out is because sales hours do not get sold. Sales is one of the departments, just as finance and facilities budgeted in the overhead of a project. Therefore their hours do not need to be estimated and therefore not included in this research. When the term white-collar department is used, it is one of the following main departments:

- Project Management,
- Engineering,
- Quality,
- Production Engineering,
- Procurement,
- Logistics.

Production control, which can be seen later in Section 2.4, is not included in this list. This exclusion is because VES restructured the floors and determined that the two parts of production control, project planning and material planning, are now part of project management and product engineering.

2.4 The bid process

The estimation of the white-collar hours is an essential part of the bid process. Figure 5 shows the steps and the order in which they occur. The process starts with registration. Registration could either be a customer that asks VES to make something or VES that participates in a tender. When more information about the project is known, VES chooses to continue forming the bid or stop because the project does not fit VES. When VES agrees that the project is suitable, they formulate a bid plan. This bid plan leads to a kick-off, where multiple project members or departments determine, together with sales, what the bid will be. During the bid review, the estimations about all three types of costs are judged. If every department approves the bid, it needs to be signed before it is handed to the customer. If the customer agrees on the bid, the last step sales does is making a production order just before they hand the project over to project management.

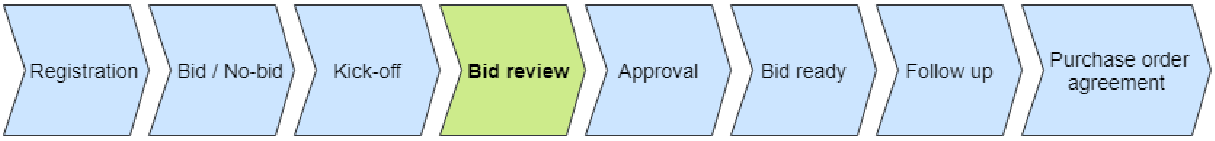
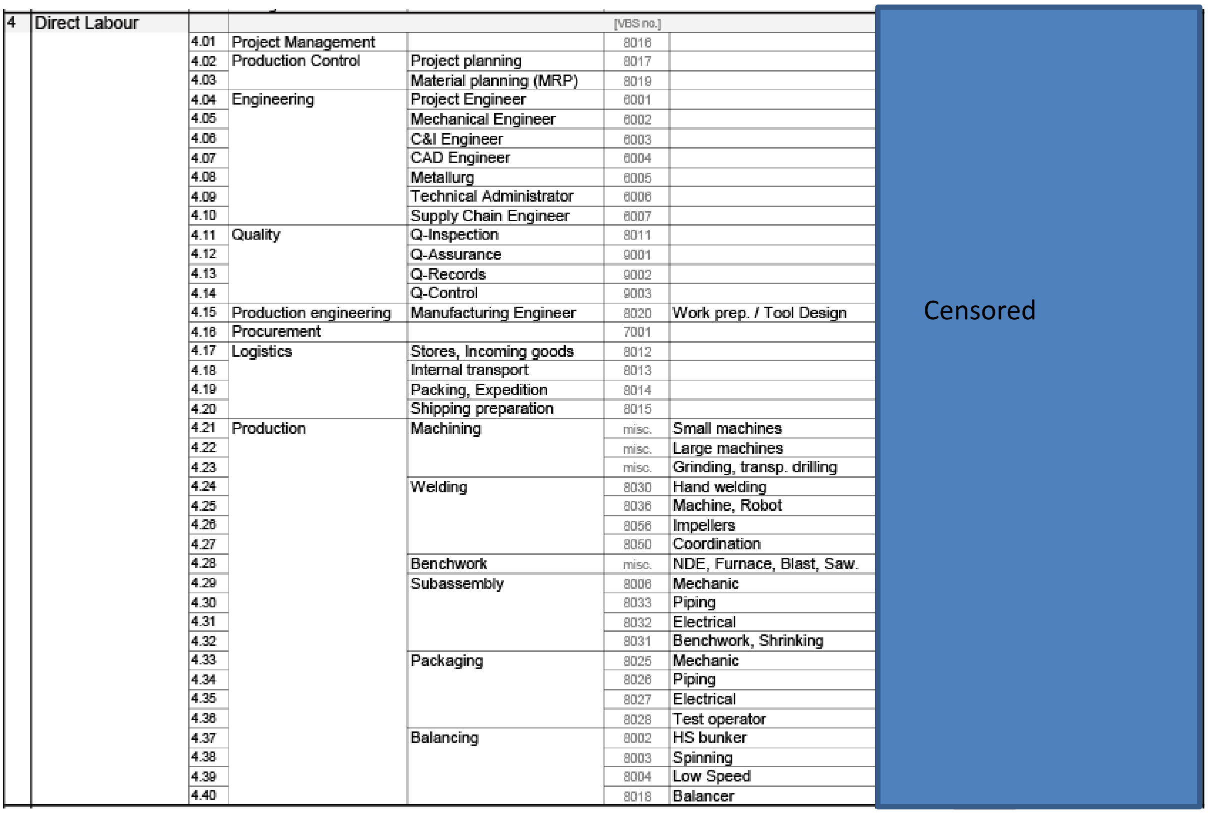


Figure 5: Bidproces in the sales phase.

The bid review is the most important part of the bid process for the research. The main reason for the importance is that after the bid review, the customer decides whether to accept the bid yes or no. So throughout the bid review, the bid is finalised. The number of hours gets locked after the department leader agreed and signed on it. What often ends up happening is that after every department made their estimations, the sales department tries to decrease the estimations where possible. It is in their interest to develop the best bid, often determined by the lowest price possible. Employees of VES call it the "kaasschaafmethode" or translated "cheese plane method". However, when the sales manager asks the department leaders if it is possible to cut some hours, they often do not indicate the cut's size or if it is even possible. They recently started a project to estimate the blue-collar hours and material costs that provide some guidance. The foundation of the estimate is the scope and specifications of the machine that needs to be packaged. In an Excel sheet, they click on the boxes corresponding with the specifications and a first estimate rolls out. However, for white-collar hours such a method is not yet realised. This research focuses on filling that gap.

Job cost sheet

When the individual departments agreed on the number of hours they want to make on a specific project, it gets summarised in a job cost sheet. A job cost sheet is a document that is part of the sales deal indicating where the customer's money goes. Figure 6 gives an example of such a sheet for a packaging project. As stated before, Production Control is no physical department, but the job cost sheet is not updated in that way.



4	Direct Labour				[VBS no.]	
		4.01	Project Management		8016	
		4.02	Production Control	Project planning	8017	
		4.03		Material planning (MRP)	8019	
		4.04	Engineering	Project Engineer	6001	
		4.05		Mechanical Engineer	6002	
		4.06		C&I Engineer	6003	
		4.07		CAD Engineer	6004	
		4.08		Metallurg	6005	
		4.09		Technical Administrator	6006	
		4.10		Supply Chain Engineer	6007	
		4.11	Quality	Q-Inspection	8011	
		4.12		Q-Assurance	9001	
		4.13		Q-Records	9002	
		4.14		Q-Control	9003	
		4.15	Production engineering	Manufacturing Engineer	8020	Work prep. / Tool Design
		4.16	Procurement		7001	
		4.17	Logistics	Stores, Incoming goods	8012	
		4.18		Internal transport	8013	
		4.19		Packing, Expedition	8014	
		4.20		Shipping preparation	8015	
		4.21	Production	Machining	misc.	Small machines
		4.22			misc.	Large machines
		4.23			misc.	Grinding, transp. drilling
		4.24		Welding	8030	Hand welding
		4.25			8036	Machine, Robot
		4.26			8056	Impellers
		4.27			8050	Coordination
		4.28		Benchwork	misc.	NDE, Furnace, Blast, Saw.
		4.29		Subassembly	8006	Mechanic
		4.30			8033	Piping
		4.31			8032	Electrical
		4.32			8031	Benchwork, Shrinking
		4.33		Packaging	8025	Mechanic
		4.34			8026	Piping
		4.35			8027	Electrical
		4.36			8028	Test operator
		4.37		Balancing	8002	HS bunker
		4.38			8003	Spinning
		4.39			8004	Low Speed
		4.40			8018	Balancer

Censored

Figure 6: Example of a job cost sheet.

The most crucial piece of information is the second last column, bordered with the red rectangle. That column states the budgeted hours per individual cost group. Groups 4.01 up to and including 4.20 are considered white-collar, while 4.21 until 4.40 are considered blue-collar. Maybe, the blue-collar cost groups say something about the white-collar hours. Logically that would make sense since white-collar hours are seen as supportive of the blue-collar hours. A correlation test will be executed later in this research to see if this assumption is valid. A discussion could be made about whether some white-collar groups are adding direct value. For example, if the machine needs to be certified, an argument could be made to say that a quality employee is needed to make the product. Hence it should be blue-collar. For the remainder of this research, every cost group up and until 4.20 will be considered white-collar.

2.5 Current estimation methods for white-collar hours for the bid review

To determine how departments estimate, all department leaders were interviewed based on the same list of questions. Since all estimations are based on expert judgement, the five phenomena are included in the questions. One of the more important and interesting questions is about the cost drivers and the comparison between projects. How do different departments see different projects, or do they see the hours factors of a project in similar ways? Table 1 shows the summary of the methods of each department.

Table 1: Departments and their main way of estimating white-collar hours for the bid review.

Department	Subdepartment	The main way of estimating
Project management	Project management	A ratio of the total amount of white-collar hours made by the other departments.
	Project planning	Half times the amount of project management hours. This method is based on expert judgement.
Logistics	Incoming goods	Based on the expected size and number of incoming shipments.
	Internal transport	Based on the expected amount of production orders, so how often material has to be moved.
	Preparation & shipping preparation	Based on the size of the machine that has to be shipped and the number of individual parts.
Engineering	-	Based on a preliminary list of actions based on the scope.
Quality	-	Based on the customer and product, a standard quality control plan is determined, which determines most of the estimation
Production Engineering	MRP & Work preparation	Based on the material order part of the scope that is known. How many product orders and the type of core products are examples.
Procurement	-	Based on extra parts of a known system, complexity, but also detailed quality requirements.

A couple of departments have specific tools to help them estimate. They are excel sheets in which, for example, a summary of the expected activities is made. The first estimate that has to be made is if a specific activity on that list is required for a new project. Based on experience, this is often relatively easy to determine. The second estimate is how long a specific activity should take. In the tool used by the engineering department, when an activity is included, it adds a standard time to the total number of hours. However, the amount of hours for that activity rarely changes in the tools. There is no feedback on those tools. This lack of feedback causes the estimate to be more like an order of magnitude estimate. A second possible problem arises when the scope defines an activity with which they have no experience. Then, a department recognises that they need to include a new activity into the estimate and include it in their tool. These estimates are likely out of the blue, based on the knowledge of the department. This research aims to provide a model to make a structured estimation.

2.6 Available data for modelling

To analyse and predict estimations, data are necessary. However, this causes multiple challenges. When the company got acquired by VDL, the ERP system changed as well. The employees went from a system called SAP to VBS (VDL Besturings Systeem). With the change, none of the old projects were copied to new drives and servers. This results in the oldest project to be from 2019. Since the project generally takes 6 to 8 months, the total number of finished packaging projects is low. On March 31 2021, a total of nine packaging projects were finished. Initially, the set-up of this research was focused on the departments' ability to follow the initial budget. However, statistical analysis with a total of nine projects is not that suitable for a thorough study. That is why the approach is chosen to start with the estimates determined at the bid phase and use the low number of finished projects to extend the model with feedback.

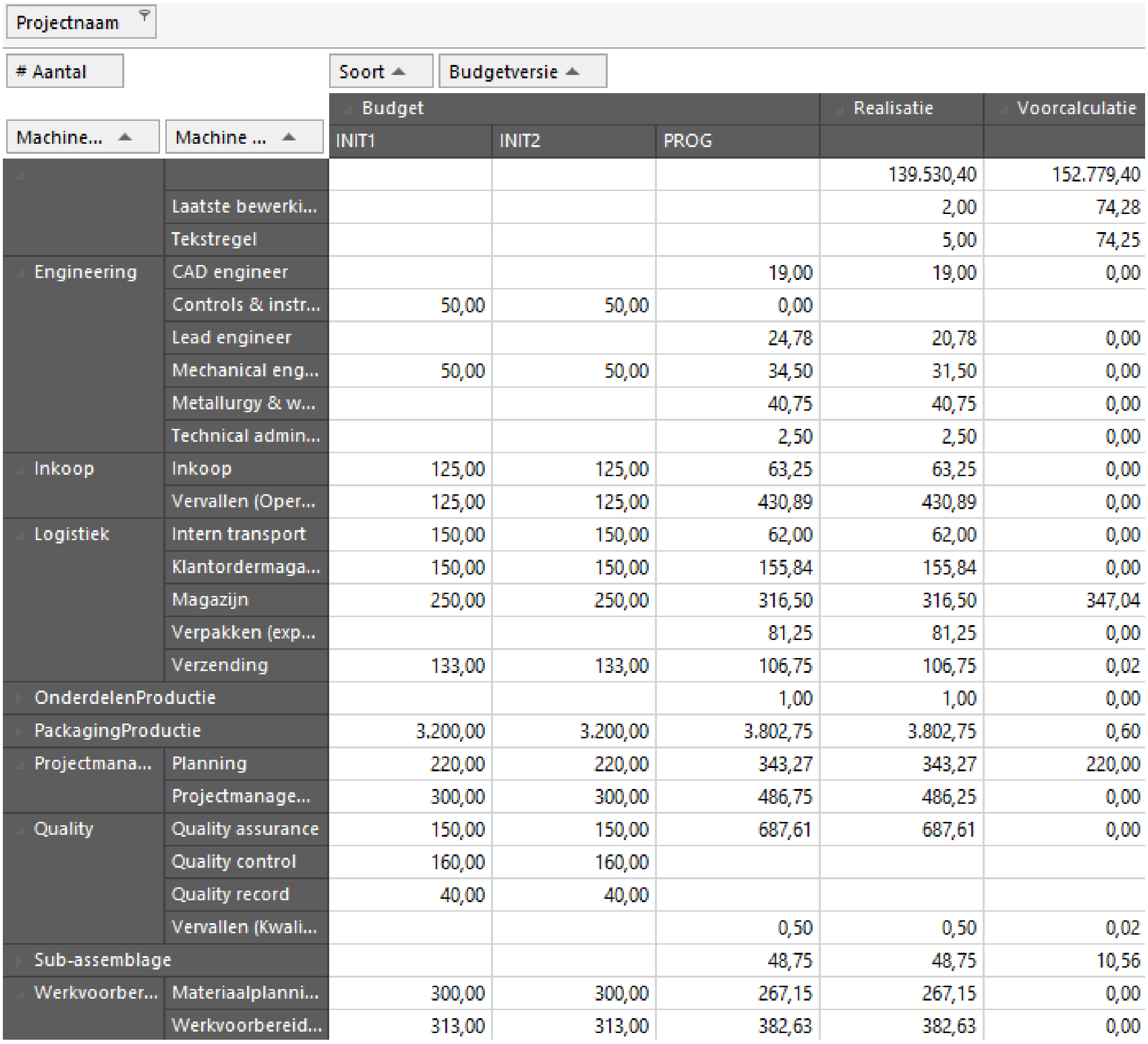
As said before, the best indication about the estimation of white-collar hours is the job cost sheet. The number of available projects with a job cost sheet is higher than nine since ongoing projects, and bounced bids are included. This brings the total relevant projects with a job cost sheet to 49.

Unfortunately, the job cost sheet does not provide all the information. Solely based on the job cost sheet, there is little or no indication about the specifications. Production hours and Engineering are the closest that could indicate the size and uniqueness of a project. VES thinks that the engineering hours have the most impact on the total number of white-collar hours. For example, working with complex materials for a known compressor that a customer requires does not influence the white-collar. However, when a new product has to be built, purchasing has to work with new materials, planners are on unknown ground, and project managers are likely to spend more time on the project. Therefore, the job cost sheet needs to be combined with some of the technical information for completeness—for example, the project's scope to sketch the entire picture.

Finished projects are the key to determining the quality of the estimation. VES keeps track of the projects in VBS. VBS has an overview of the finances, but also hours, of all projects. This overview, exampled by Figure 7, contains a couple of columns with information. The columns are:

- INIT1: The originally sold number of hours.
- INIT2: The revised sold number of hours if the customer agreed. A change in scope can cause this.
- PROG: The prognosis from the project controllers, the estimate what the total number will be. Not relevant for this research.
- Realisation: The registered or booked number of hours. This increases while a project is ongoing and indicates the final number of hours when the project is finished.
- Pre-calculation: Shows the planned hours for part production and sub-assembly determined by work preparation. Not relevant for this research.

The rows in the systems are the same as the cost groups in Figure 6. Therefore, the comparison can be made between INIT1 or INIT2 and the realisation. INIT1 is used when the scope did not change during a project, and if the scope did change, INIT2 is used. The difference between the INIT and the realisation shows the number of hours that a cost group worked more or less on that specific project.



Projectnaam

Aantal

Soort ▲ Budgetversie ▲

Machine... ▲	Machine ... ▲	Budget			Realisatie	Voorcalculatie
		INIT1	INIT2	PROG		
					139.530,40	152.779,40
	Laatste bewerki...				2,00	74,28
	Tekstregel				5,00	74,25
Engineering	CAD engineer			19,00	19,00	0,00
	Controls & instr...	50,00	50,00	0,00		
	Lead engineer			24,78	20,78	0,00
	Mechanical eng...	50,00	50,00	34,50	31,50	0,00
	Metallurgy & w...			40,75	40,75	0,00
	Technical admin...			2,50	2,50	0,00
Inkoop	Inkoop	125,00	125,00	63,25	63,25	0,00
	Vervallen (Oper...	125,00	125,00	430,89	430,89	0,00
Logistiek	Intern transport	150,00	150,00	62,00	62,00	0,00
	Klantordermaga...	150,00	150,00	155,84	155,84	0,00
	Magazijn	250,00	250,00	316,50	316,50	347,04
	Verpakken (exp...			81,25	81,25	0,00
	Verzending	133,00	133,00	106,75	106,75	0,02
OnderdelenProductie				1,00	1,00	0,00
PackagingProductie		3.200,00	3.200,00	3.802,75	3.802,75	0,60
Projectmana...	Planning	220,00	220,00	343,27	343,27	220,00
	Projectmanage...	300,00	300,00	486,75	486,25	0,00
Quality	Quality assurance	150,00	150,00	687,61	687,61	0,00
	Quality control	160,00	160,00			
	Quality record	40,00	40,00			
	Vervallen (Kwali...			0,50	0,50	0,02
Sub-assemblage				48,75	48,75	10,56
Werkvoorber...	Materiaalplanni...	300,00	300,00	267,15	267,15	0,00
	Werkvoorbereid...	313,00	313,00	382,63	382,63	0,00

Figure 7: Example of data of a finished project

This difference gives a first indication of how the company did, per cost group. However, it only gives a numerical value and does not tell if, for example, the budget overrun was caused by a specific cost group, the customer, or even a supplier. However, including the cause of an overrun might be hard to model. VES tries to evaluate projects with a lessons learned document, but the quality of the feedback differs per project. The possibility of using this information is kept in mind throughout the literature review.

Conclusion

The white-collar hours are often estimated with the help of a small excel tool. The next chapters will focus on how to use the 49 during the bid phase estimated projects and the finished projects to made data-driven estimations. The main challenge will be the small amount of data, since the total number of projects is not much.

3. Literature review

The literature review consists of four parts. Section 3.1 introduces statistical learning. The introduction provides a basis to answer some of the research questions. Section 3.2 considers the different types of models or methods, both for the main model and the extended model, that are suitable regarding the limitations of this research. Section 3.4 explains the different ways of validating and testing, focusing on small data since this makes validating and testing harder. Finally, Section 3.4 concludes the chapter and gives a guideline for the techniques used onwards.

According to (Gonfalonieri, 2019) two common approaches can help build predictive models from small data sets. The first approach is to use a simpler classifier model and the second approach is to use ensemble methods, in which voting between classifiers can compensate for individual over-learning. Examples of models that should be suitable are naïve Bayes methods, linear models and decision trees. These are focused on making the small amount of data work.

Two books, "The Elements of Statistical Learning" (Hastie, Tibshirani, & Friedman, 2008) and "Introduction to Statistical Learning with Applications in R" (James, Witten, Hastie, & Tibshirani, 2017) provide a clear starting point. The main advantage of the second book is the help of how to use the data to make models. The use of the second book also implicates that R & RStudio are the software of choice.

3.1 Statistical learning

As stated in Section 1.6, the main modelling style will be statistical learning. However, what is statistical learning? The definition from the website deepai.org describes it:

"*Statistical learning theory is the broad framework for studying the concept of inference in both supervised and unsupervised machine learning. Inference covers the entire spectrum of machine learning, from gaining knowledge, making predictions or decisions and constructing models from a set of labelled or unlabelled data. The entire process is stated in a statistical framework, with every assumption stated mathematically as a null or alternative hypothesis.*" (Statistical Learning Theory, sd)

Within the statistical learning framework, a couple of subcategories are known. To know what tools suit the research the best, these subcategories need to be known. A complication of using statistical learning is that a model needs to be trained and tested. This generally means that the initial dataset is split in two. For example, the first part is needed to determine the values of specific parameters for the model. The second dataset is then used to test the model based on dataset one. If the model uses all the data simultaneously, the model will overfit. This means that the model cannot think outside of the box. This difference is usually shown in a training error and a test error. The training error is the error when the dataset that is used to determine the parameters is fed back into the model. The test error is the error when the test dataset is fed to the model. Both errors are types of statistical tests. Within statistics, a more extensive dataset means better statistical soundness. However, VES does not provide a big data set. Therefore, it is essential to consider methods that can provide sufficient train and test errors, even though the dataset is not as large as an analyst would typically expect when doing statistical analysis.

Supervised learning versus unsupervised learning

The difference between these two categories is the modelling approach. In supervised learning, the inputs or features provide an estimate of a value as an output. This output can be used to tell the model how well it did to predict a specific value. An example of supervised learning is predicting what the temperature will be for the next day. A prediction of 33 degrees Celsius would be logical for a Mediterranean region in summer, but not for a winter in the north of Europe. Since the learning is supervised, the model will learn. In this case, depending on region and season.

On the other hand, unsupervised learning does not have an output value. Generally, unsupervised learning aims to search through the data for links, groups or clusters. Throughout this process, there is no interference of judgement. Since in this research finished projects are available to test estimates, the choice to use supervised learning is an easy one. Unsupervised would not make sense in this specific situation.

Regression problem versus classification problem

The weather problem from the last paragraph is a perfect example of a regression problem. In a regression problem, an analyst is working with continuous values. For example, a classification problem judges if the weather is good enough to barbecue. The answer to that question is yes or no. So depending on what the output needs to be, a decision is made on how the output should look. In both cases, the same features can be used. In a supervised learning model, season and region can predict the next day's temperature for barbecue. In an unsupervised learning environment, regression is not possible.

3.2 Different types of models

This section covers three different types of models: multiple linear regression, decision trees, and random forest. The last two are similar in some ways but can come up with totally different outputs and are therefore considered separately. Many more methods could be used, but the focus is on these three because of their straightforwardness. Each section explains the basics per method and gives examples of how the methods are used in similar cases in recent years. Unfortunately, examples out of the oil and gas industry are rare. So the reference needs to come from another industry. Industries that are similar and more widely studied are civil engineering and project cost estimation of software projects. The similarity is based on the supportive activities needed to carry out the main task.

3.2.1 Multiple linear regression

Multiple linear regression or MLR is an extension of linear regression. The goal of the model is to take an input vector $X^T = (X_1, X_2, \dots, X_p)$ and want to predict a real-valued output Y (Hastie, Tibshirani, & Friedman, 2008). T stands for the instance or situation, and p is the number of variables. Equation 1 gives the form of the model.

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \quad (1)$$

In this equation, X_j are the variables, and β_j are the unknown parameters. A multiple linear regression model means that p is greater than 1. Multiple variables impact the outcome, which is f(X). When the jth value out of the vector X^T is not impactful enough, the model will set β_j to zero. The main assumption, logically, is that the features and output behave in some linear way. The equation shows the most basic MLR model, but it is the basis. In an extended MLR, it is possible to work with multiple outputs. In that scenario, each type of output Y has its linear model.

Applications in literature

A simple example of how multiple linear regression can predict project characteristics is given by Fandopa, Alisjahbana & Ma'soem (2020). In their paper try to increase the accuracy in estimating the cost of the borrowed land. They define 10 features, or as they call it, feature variables to estimate the dependent variable cost of the borrowed landfill work in Rupiah per m^3. They also show that with 20 projects as input, the R^2 error can still be significant at 0.968. However, this paper does not fully connect to this research because no training is included. They tried to fit a multiple linear model as good as possible on their data, so it is likely overfitted to the real world.

Araya et al. (2020) provide a similar example. They took data from 38 bridge replacement projects to predict the engineering man-hours. The main thing they show is the ease in which a low number of projects, together with all different types of variables, binary, categorical, and continuous, can still lead to a model that predicts well. Nevertheless, again, they only made a model that fit the data as well as possible, leaving no room for training.

The last example comes from Persad, O'Connor & Varghese (1995). This older research presents statistical models for forecasting the engineering manpower requirements for highway preconstruction activities. They use a more extensive dataset with a total of 758 projects. Their models show the application of having one of two feature variables. Their first model only uses construction costs as an indicator for the manpower requirements, while their second model use construction costs and project type. They show that two features predict a lot better than one and even claim that the second model is an excellent predictor. However, the highest obtained R^2 is 0.6761, which seems to leave some room for improvement.

3.2.2 Decision trees

Decision trees or tree-based methods partition the feature space into a set of rectangles and then fit a simple model in each one (Hastie, Tibshirani, & Friedman, 2008). Tree-based methods use splits in the data to make "branches" Figure 8 gives an example in which only two features impact the outcome Y.

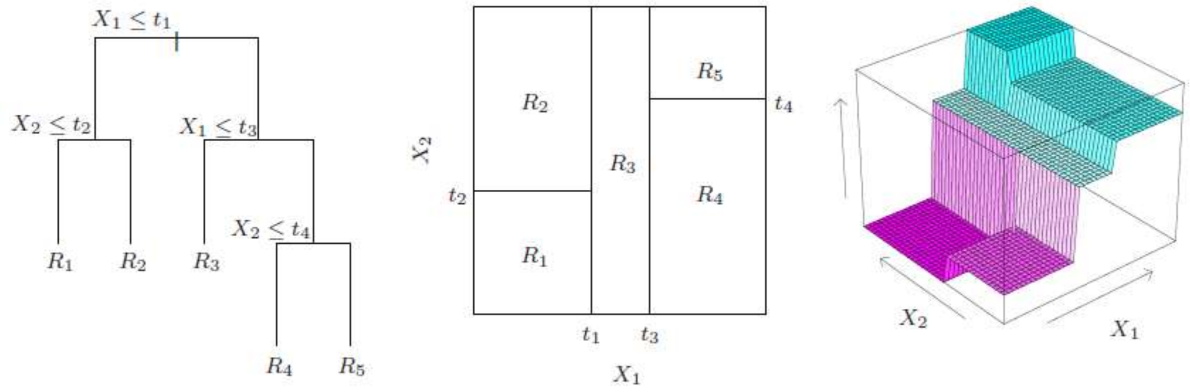


Figure 8: Visualisations for the tree-based method. Copied from Hastie, Tibshirani & Friedmann, 2008, p. 306.

The left diagram represents the order of decisions you have to make to end up in one of the five regions. The middle diagram then shows the area's corresponding to those decisions. The right diagram indicates that every area has its Y value, so a different prediction. Equation 2 shows the formula to predict Y in this example.

$$\hat{f}(X) = \sum_{m=1}^{5} c_m I\{(X_1, X_2) \in R_m\} \ (2)$$

This formula only works after the model is made, just like the formula of MLR. Based on the example, the constant c_m and region R_m get determined by the model. Function I determines which region is active for a specific X_1 and X_2. It sets the value for that region to one and all the other to zero. Therefore, the summation only uses the one constant that corresponds with the region, which is the prediction.

An algorithm makes the tree. The algorithm decides on the splitting variables, split points and the shape of the tree. An important decision that has to be made while using this method is the size of the tree. If the tree has hundreds of branches, it is likely to overfit, while a small tree might be too general. That is why tree size is a tuning parameter. Another essential property of the method is how the tree is built step by step. The algorithm in the appendix uses an approach that chooses the best split in every iteration of the algorithm. Therefore it is a greedy algorithm. A property of greedy algorithms is that they cannot guarantee to find the best solution in the solutions space. For decision trees, this property means that the algorithm does not find the best tree with the best splits, to lower errors or improve the prediction.

Pruning and boosting

Multiple methods exist to increase the predictive power of a tree-based method. A well-known method is pruning. When you prune a tree, you get rid of branches that have no predicting power. Pruning gets rid of statistical noise, lowers the size because it prunes the branches and generally improves the accuracy of your tree.

Boosting is a method, often in the form of an algorithm, that seeks to improve the prediction power by training a sequence of weak models, each compensating the weaknesses of its predecessors (Zhang, 2016). The starting point of a boosting algorithm is a weak model, which could be a regression or a shallow decision tree, and improves it. Zhang introduces two algorithms: Adaptive Boosting (AdaBoost) and Gradient Boosting. AdaBoost is mainly for classification problems, while Gradient Boosting can be used for both classification and regression.

3.2.3 Random forest

Random forest is a method that uses the tree-based method together with a method called bootstrap aggregation (bagging). Bagging is a technique for reducing the variance of an estimated prediction function (Hastie, Tibshirani, & Friedman, 2008). This technique is done by testing a regression tree to multiple bootstrap samples of the training data and average the result. The idea behind the bootstrap is explained in Section 3.3.2. In the case of random forest, it means that the features considered per split are randomised. Thus, the random forest method creates multiple de-correlated trees on the training data. Afterwards, the method averages all these trees (the forest) to provide a prediction model. Hastie, Tibshirani & Friedman provided an algorithm in their book.

Applications in literature

One of the first significant mentions in literature is by Breidman (2001). He studied different ways of using the method on different datasets and concluded that random forest is an effective tool in prediction. He states that injecting the right kind of randomness makes random forest accurate classifiers and regressors. Furthermore, the framework in terms of the strength of the individual predictors and their correlation gives insight into the ability of the random forest to predict. He also concludes that random inputs and random features are more suitable for classification than for regression.

Awada, Srour, & Srour (2021) have recently used random forest to forecast delays during a project. Their method uses concrete pouring requests as an example of a site data stream to predict if the request is accepted or not. They divided their 933 instances randomly into a train and a test set by a 75-25 ratio. The result was an accuracy of 91%, but in their conclusion, they address one crucial flaw. The model is built on past site data and thus will inherently include the past behaviour of contractors. Unfortunately, they do suggest when. It seems inefficient to analyse when ten more instances are in the dataset, but this is a more general concern.

3.2.4 Comparisons

A common subject in papers is the comparison between methods. The best method differs per situation and dataset. An example of a study that considers multiple methods is Pahno, Yang & Kim (2021). They compare MLR, regression tree, random forest, and a method called extreme gradient boosting (XGBoost), a type of boosting. With their dataset about the subgrade resilient modulus of sand, they compare the R2 values and the root mean square error (RMSE) and the importance of variables per method. They conclude that the variation in variable importance and ranking across models is likely due to the difference in model structures and algorithms used for model training or fitting.

Elmousalami (2021) presents a comparison of 20 machine learning models. This comparison is based on a dataset of Egyptian field canal improvement projects to model the conceptual costs. Again, XGBoost ended up as the most accurate, based on the adjusted R^2 and the mean absolute percentage error (MAPE). They explain the importance of ensemble methods for improving accuracy. However, ensemble methods have a limitation in which it becomes harder to interpret the results. This is caused by the fact that ensemble methods use, per definition, multiple learning algorithms at the same time.

Varshini, Kumari & Varadarajan (2021) provide a recent example of predicting the software project estimation. Software project estimation consists of time estimation, resource estimation, cost estimation, and effort estimation. Effort estimation focuses on predicting the number of hours of work to develop software. For determining the best estimation method, they compared single model approaches and ensemble model approaches. They used publicly available datasets and judged the performance based on the mean absolute error, RMSE, and R^2. They conclude that the stacking random forest performed the best compared to single models in their case. Stacking often considered heterogeneous weak learners, learns them in parallel and combines them by training a meta-model to output a prediction based on the different weak model predictions (Rocca, 2019).

Another paper from Varshini, together with Kumari, Janani, & Soundariya (2021), focuses on analogy based estimations, regression estimations, classification approaches, and deep learning algorithms to predict the software effort estimation. The methods deepnet, neuralnet, support vector machine and random forest were input for the comparison. The evaluation metrics are MAE, RMSE. MSE and R^2. With these metrics, they found that the random forest outperforms the other methods. Some datasets are used in both Varshini papers, but they do not address if that might cause the random forest method to be superior in both papers.

3.3 Testing and validating

After the choice of method, the method needs to be tested and validated. As stated before, a good statistical model finds the balance between a test error and a training error. Therefore, this section focuses on the evaluation metrics and ways to improve those working with small data.

3.3.1 Evaluation metrics

Evaluation metrics help to judge the accuracy and quality of the model. In the most general idea, the metrics compare the original dataset and the dataset made by the model. The error per data point is then the difference between both values for a specific input. Varshini, Kumari, Janani, & Soundariya (2021) give a quick summary of the metrics.

- Mean absolute error: the average sum of all absolute errors.
- Mean squared error: the average of square errors.
- Root mean squared error: the square root of the mean squared error, or the average of standard deviations of evaluated deviation.
- R^2: Proportion of the variance in the dependent variable that is predictable from the feature variable(s). Also called the co-efficient of determination.

One metric that Varshini et al. (2021) do not consider is the mean absolute percentage error, which provides a ratio on how well the forecasted value fits the actual value. Of course, every metric has some value in judging the quality of the model, but most evaluations start with the MSE or R^2.

Quantile regression forests

A specific type of evaluation for a random forest is Quantile Regression Forests (QRF). This method, first mentioned by Meinshausen in 2006, provides more than an estimate when predicting with random forests. Throughout the algorithm, it tracks all possible outcomes summarised in a dataset. This dataset then provides the opportunity for extra analysis, such as the determination of certain quantiles.

3.3.2 Possibilities with small data

In an ideal modelling situation, lots of data are available. In real life, however, this is not always the case. Researchers already concluded that years ago and started to think of ways to get the maximum out of small data sets. This section explains two different methods of handling small data. One method is to generate more data, and the other method is to train and test smartly with the data on hand.

Generating data

Another approach is to generate data. One of the ways to end up with more data is to augment your data. (Rothmann, 2019) states that augmentation techniques allow you to produce many more "semi-unique" data points for training your model. An example of augmenting data is adding Gaussian (normal distributed) noise.

The last approach is to use synthetic data. Both Gonfaloniere and Rothmann describe this possibility. It means creating fake data. One of the techniques is the Synthetic Minority Over-sampling Technique (SMOTE). Smote takes the minority class data points and creates new data points between any two nearest data points joined by a straight line (Gonfalonieri, 2019).

Testing approaches with small data

Different testing methods exist that help to get the most value out of the data. Innovative solutions on how to divide your train and test data are necessary when dealing with small data. Two well-known methods are K-fold cross-validation and the bootstrap method.

K-fold cross-validation

This is an error finding method, especially for scarce data. To work around this obstacle, K-fold cross-validation uses part of the available data to fit the model and a different part to test it (Hastie, Tibshirani, & Friedman, 2008). First, the data is split into K roughly equal-sized parts. Then, for the k^th^ part, for example, the third, the model is fit to all the other K-1 parts and calculate the prediction error of the fitted model when predicting the k^th^ part of the data. This is done for all parts, and the K estimates of the prediction error are combined.

Bootstrap

The basic idea of the bootstrap is to randomly draw datasets with replacement from training data, each sample the same size as the original training set. This is done B times, producing bootstrap datasets (Hastie, Tibshirani, & Friedman, 2008). B stands for the number of new, bootstrapped datasets. Table 2 provides an example with B is 3. An example, in this case, is the number of finished packaging projects per quartile.

Table 2: Example of a bootstrap.

Finished projects	Original	Bootstrap 1	Bootstrap 2	Bootstrap 3
Quartile 1	1	1	2	2
Quartile 2	2	1	2	3
Quartile 3	3	3	4	3
Quartile 4	4	4	4	3

After those bootstrap data sets are made, the already existing model is refitted. This way, the bootstrap method tries to get a more realistic grasp of the model's errors.

3.4 Conclusion

The chapter starts with an explanation of statistical learning and its types. The most suitable type for this research is a supervised regression problem. The most usable output is the number of hours per department. Within this part of statistical learning, plenty of methods exist that can predict. The research questions are the basis of the conclusion of this chapter. Each of the following sections provides the answer to continue the research on.

What are the best indicators to model with, based on experience and data?

The methods determine themselves what the best indicators are. There is no clear answer to this question before modelling. Most models even show the effectiveness of a specific parameter in their summary statistics.

What model is the most suitable regarding small amounts of data?

The review addresses three models that are all suitable regarding small data. Each model has its complications, but methods exist to get the most out of a small data set. That can either be a model working on artificial data or validating the model with specific tricks to get the best model possible.

What is the best method for the main model to estimate the white-collar hours?

The three models have advantages and disadvantages. Table 3 provides a quick summary.

Table 3: Advantages and disadvantages of different modelling options.

	Multiple linear regression	Decision trees	Random forest
Advantages:	Easy to interpret	Easy to interpret	Powerful and accurate
	Multiple outcomes	Extremely fast	Low chance of overfitting
		You can show how the decisions are made, which makes the method easily reproducible	
Disadvantages:	Restricted to linearity	Decision trees are likely to overfit the data	Relatively slow model
		Requirement of algorithms that determine the optimal choice at each node	It cannot be used for linear methods
			Difficult to use for high dimensional data

The summary and the literature do not answer this question. It is highly dependent on the dataset but also on how the model shows the output. In the end, we determined that all three methods, multiple linear regression, decision tree and the random forest are the models of choice. Since it is highly dependent on the dataset if a method is suitable for estimating we choose to use all three. Theoretically, even more methods could be used to estimate because these three are not the only ones in the field of statistical learning, but a choice has to be made and these methods shine in simplicity, which suits small data. When comparing the three methods, interpretability is the most crucial characteristic which decision trees guarantee, but the accuracy of random forests and multiple linear regression might overrule the interpretability.

4. Solution design

This chapter contains the data preparation, the steps taken for both methods and some explanation about the evaluation metrics. Section 4.1 consists of the data preparation. Preparation includes cleaning the data and determining the features. Section 4.2 shows the three methods. Three flow charts show the steps per method to determine a predictive model. Section 4.3 shows the characteristics of these models so they become comparable. Chapter 5 gives the values based on the current dataset.

4.1 Preparation

The starting point is the projects under the category VES-Packaging or Siemens-Packaging. The second filter is the availability of the estimates. Based on information from the ERP system and the file server, a total of 49 almost complete projects are compatible. Unfortunately, the order cost sheets on the server have two different formats. 18 are in the old format and 31 are in the new format. The OCS shown in Chapter 2 is the new format that VES uses currently. The formats do differ severely. The old one has another way of determining blue-collar hours, different types of work that are part of the department called planning & control, and no specification for the engineering hours. A couple of changes are necessary to combine the two. First, all the cost groups for blue-collar hours form one large feature variable. Second, the three planning cost groups from the new format make a sum of all planning variables. The disadvantage of this change is that it follows the old format. An analysis is still possible for the individual planning cost groups but with fewer projects. Less usable projects will likely influence the prediction power, but the sum of the variables gives some indication. The last change is the summation of all the engineering, quality and logistic cost groups. The summation of the white-collar cost groups is preferable for small data. However, the old format forces it, which means the individual cost groups will play no role in the research.

Response variables

Figure 6 shows all the individual cost groups. Out of those 40 cost groups, 10 variables serve as input for the model. To summarise:

- Project Management (PM), group 4.01;
- Project Planning (PP), group 4.02;
- Material Planning (MRP), group 4.03;
- Engineering (Eng), groups 4.04 to 4.10;
- Quality (Q), groups 4.11 to 4.14;
- Production Engineering (WVB), group 4.15;
- Procurement (Proc), group 4.16;
- Logistics (L), groups 4.17 to 4.20;
- Production (SUM_BC), groups 4.21 to 4.40.

The main department production control is broken up, because of the formatting in the old format. That format only had the department planning, which contained PP, MRP and WVB. Therefore the number of projects in the dataset with those three values is lower than the number with only planning. **From now on the different response variables listed above will be addressed as departments.**

Feature variables

However, the OCS's only provide one feature variable, the sum of blue-collar hours, next to all the response variables. One feature variable can be enough in some cases, but not in this one. The information about a project in the ERP and file server heavily determined what features variables were chosen. It was hard to find specific information, so the feature variables are forced to be generic. After discussing the possibilities with VES, a total of seven different feature variables will be tested on their ability to predict white-collar hours. Table 4 gives a summary of the feature variables.

Table 4: Overview of the feature variables.

Name	Definition	Type	Values
Sum of BC	Number of predicted hours for all blue-collar activities	Numerical	-
Category Customer	Category of the customer. Experience of working with a customer might influence the estimations.	Categorical	CUST_X, CUST_Y, Other
Main Component	The main component of the packaging project.	Categorical	Turbine, Compressor, Both, Test
Pieces	Number of machines	Numerical	-
Design	The number of different designs/types needed. So the number of designs can never be higher than the number of pieces.	Numerical	-
Specifications	This variable is an indication of how demanding an end-customer is. The direct customer and end customer are often different companies. This variable indicates the demand of the end-customer for the complexity of the packaging project.	Categorical	Low, Medium, High
Onshore	If the machine(s) are built for an onshore or offshore environment	Binary	True – Onshore False - Offshore

The use of binary and numerical variables is straightforward. However, the odd one out is the categorical variable. This is because R needs extra steps to deal with categorical variables e.g. the category customer variable. R split this variable into a binary variable system to be able to make a decision tree. To do so, R makes a table like Table 5.

Table 5: From categorical to binary in R.

	Siemens	Other
CUST_X	0	0
CUST_Y	1	0
Other	0	1

This way, a decision tree can branch based on one of these values e.g. if it branches on, for example, CUST_Y the other branch contains CUST_X and Other values. This looks like one-hot coding, but is not since the first row contains a double zero.

With the feature variables added to the data, data are ready to analyse. If at any moment in time VES wants to expand the analysis, a new project, feature variable or response variable can easily be added in the excel document. Appendix B shows the style of the document and its contents.

Cleaning

After the importation of the excel document in R, the dataset needs to be cleaned from missing values, called NA values. NA values disturb the analysis and multiple formulas in R do not work when the data contain NA values. After the data gathering, a lot of NA values were in the feature variable onshore. Therefore it is left out of this analysis because the lack of information meant that too many projects had the value NA. This lowered the number of usable projects significantly which is far from ideal.

4.2 Methods

Chapter 3 concludes that we select multiple linear regression, decision trees, and random forests as methods to implement. All methods need slightly different approaches but the goal is the same for all methods; to use the same dataset to estimate the number of hours needed per project per response variable.

Multiple Linear Regression

The first method is MLR. To recap, an MLR model gives every feature variable one value to estimate the response variable. In the case of a categorical variable, R automatically splits it up like explained in Table 5. Only considering the feature variables SUM_BC and main component, the model for the department PM looks like this:

$$\begin{aligned} &Number\ PM\ hours \\ &\qquad = a * SUM_{BC} + b * MainComponent_{Turbine} + c * MainComponent_{Both} + d \\ &\qquad * MainComponent_{Other} + e \end{aligned}$$

If all the feature variables would be used, the equation would be filled with more feature variables and there corresponding constant, which is the letter in front of it. The software of choice determines the values a to e. The value e is also called the intercept. Making a model per department requires a couple of steps. First, we should determine which feature variables are significant enough to be allowed in the model. This is called model selection. Two examples of such a selection are forwards and backwards. Forwards selection starts with one variable, makes the best fit possible, and then ups the number of variables with one. This is repeated until all variables are used and the best configuration is chosen. Backwards selection starts with all variables and deletes the least significant one. The last way of model selection is using a validation set. This provides a way to train and test with the dataset instead of the final training and testing of the estimates. The data is split into a train and a test set, where the training dataset and a selection e.g. forwards determine the model. This model is tested against the test set and indicates the test error per number of variables included. Figure 9 gives an example of a graph showing the train and test error. The x-axis shows the index. This index is the number of feature variables used to make a model to then test on. The reason it goes up to ten, and not to 6 are the categorical variables. With the split explained in Table 5, the category customer has two individual variables, specifications has two, and main component has three, since the double zero variable is not used in the formulas by R. This means the total number of variables R can choose from is 10, instead of 6. The black line being lowest in the position of index 1 means that only one feature variable should be used modelling that department.

This study uses this method to select the number of feature variables. The next step is to plug the feature variables in the code and a model to estimate the number of hours for a certain department comes out. The final model is based on the entire dataset, since the training and testing and therefore the chance of overfitting is moved a step upwards. The last step is to calculate the (train) MSE.

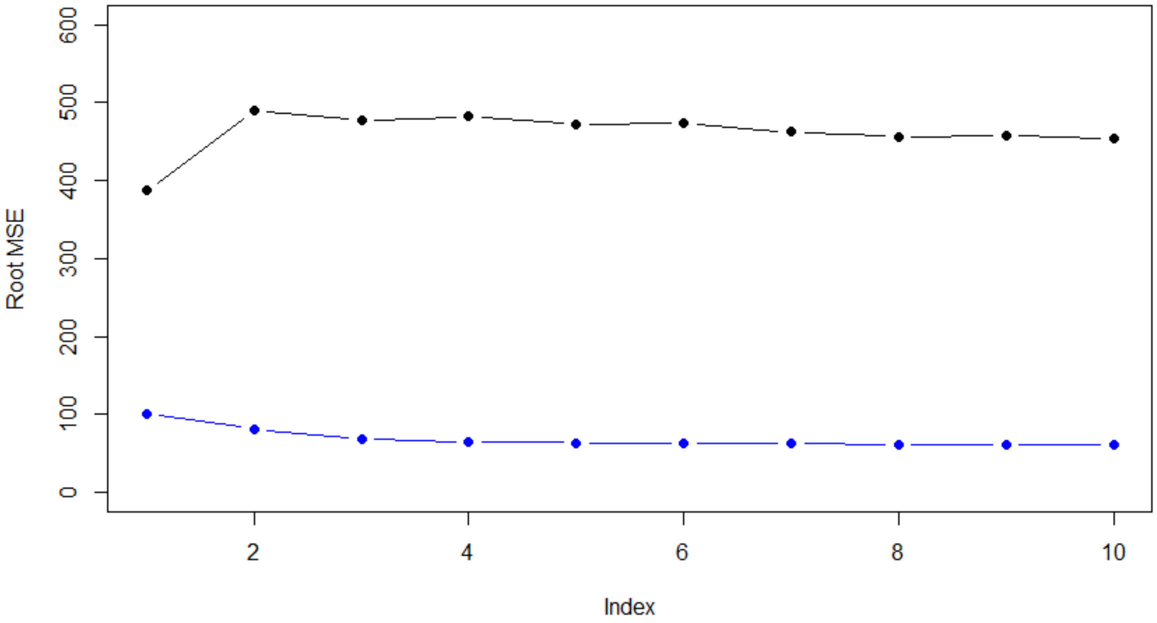


Figure 9: Model selection for MLR. The blue line is the training error and the black line is the test error.

Decision tree

The method decision tree has different steps than MLR, logically. The biggest difference is that the determination of the number of terminal nodes (endpoints) instead of the feature selection is the first step. The algorithm behind the building of the tree determines per split what the best feature variable is to split on, so if a variable is not significant it will never be chosen. The decision tree method starts with running the algorithm with the entire dataset. Next, that specific tree is cross-validated to check if it is possible to get better results with fewer endpoints. If this is true, the tree is pruned with the best number of endpoints provided as an input for the pruning algorithm. Figure 10 shows an example of the output, given the cross-validation of a tree, of a function in R. As can be seen, the initial number of endpoints is 6, but the performance is improved when the tree has 2. So therefore the decision tree model needs to be pruned to ensure the best results.

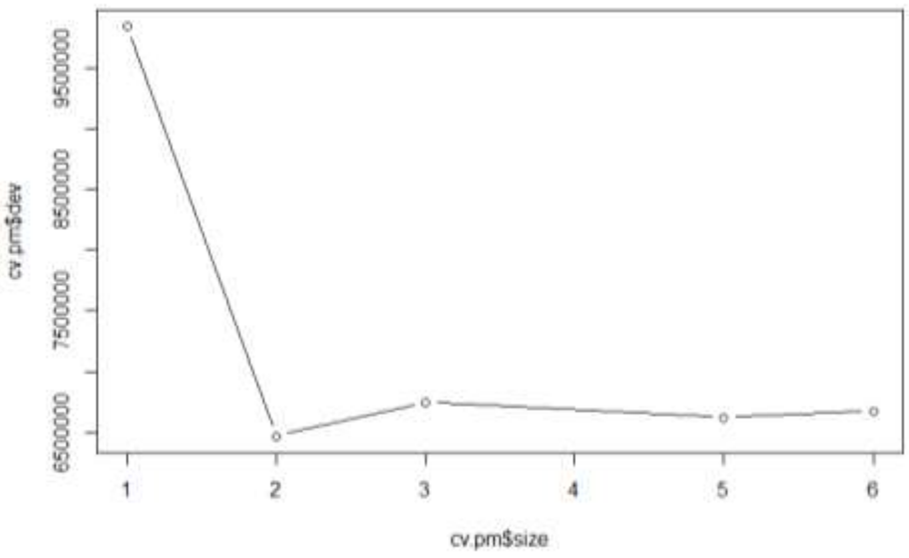


Figure 10: Cross-validation graph.

Using a train and test set to determine the best model is hard with small data. When you randomly divide the projects and use 70% to train, the test results are heavily dependent on the choice of project. For example, if the five biggest projects are all in the test set, the training is done without them and influence the test error negatively. Making multiple trees and using averaging ends up in a random forest so that is not a possibility. Therefore, the decision tree model contains all the projects in the data set. However, training and testing still have a purpose in the decision tree model. The MSE provides insight after making the model, but that MSE is based on some sort of overfitting. To give a more reliable indication of the adaptability of the model, a method called K-folds cross-validation expands the analysis. This method divides the dataset into a number of (k) folds. Then the mean of all the test MSEs of every iteration indicates how well the model does when having to estimate a new project. The number of folds is set to 5, which makes the training set 80% and the test set 20%.

The last part of the decision tree analysis is to "determine the quantiles per terminal node" step. This is an extra step designed to provide an extra layer of information about the quality of the model. It serves as a second way to determine the error, by looking at the all the values in a terminal node. The original method is to compare them with the initial value and determine the (R)MSE. This method provides more insight on how that mean is constructed. The determination works as follows:

1. Find the unique terminal nodes;
2. Follow the tree per project to find its corresponding terminal node;
3. Save the known number of hours to an array with all the projects that end up in that terminal node;
4. Determine the quantiles per dataset based on all the projects in the terminal node.
5.

Table 6 shows the quantiles for the department PM.

Table 6: Example of the intervals.

Quantile	Terminal node 1	Terminal node 2	Terminal node 3
0%	0	200	940
25%	202,5	442,5	1220,5
50%	350	650	1549
75%	478	995	1600
100%	1063	1464	1616

The 0% means the lowest value of the real number of hours in that terminal node, while 100% is the highest. The main takeaway is the overlap of the intervals. In this scenario, the 25% to 75% intervals overlap the slightest between nodes 1 and 2, where 0% to 100% has a lot of overlap. This means that a true estimate of for example 700 could have end up in either node, with their corresponding estimate every node has in a decision tree. Assuming the 50% value, 350 and 650, is the value connected to the node and that the correct estimate of a project would be 700, the project could end up in either node, while the second node causes the MSE to be lower. A lot of overlap means a high risk for a project to walk the wrong path in the three, increasing the MSE. No overlap indicates that an average project is likely to end up in the "correct" node, while an outlier project would likely end up in the "wrong" node.

Figure 11 summarises the steps required to model.

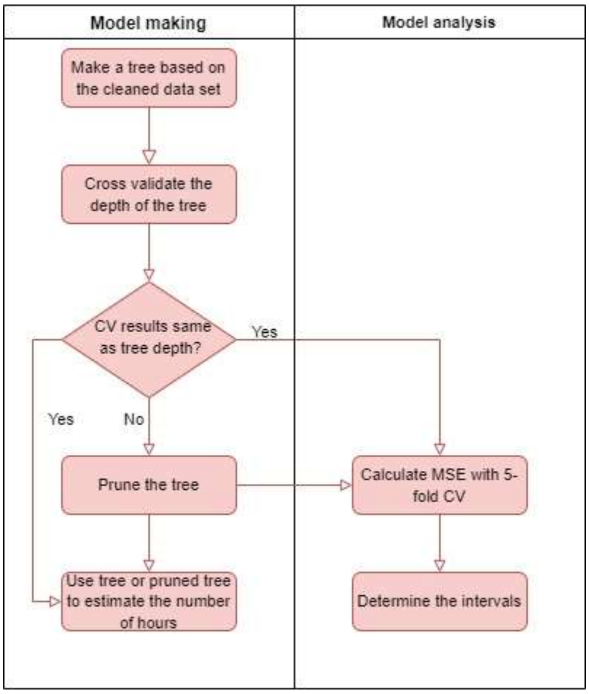

Model making
Model analysis
Make a tree based on the cleaned data set
Cross validate the depth of the tree
CV results same as tree depth?
Yes
Yes
No
Prune the tree
Use tree or pruned tree to estimate the number of hours
Calculate MSE with 5-fold CV
Determine the intervals

Figure 11: Summary of the modelling steps for DT.

Random forest

The last method is the random forest. The first step is to divide the data. The first step is to divide the data. This analysis bases its training and testing on a 70-30 ratio. This is a quite common ratio for small datasets (Tokuç, 2021). In random forests, there is no need for cross-validation or a separate test set to get an unbiased estimate of the test set error. It is estimated internally, during the run (Breiman & Cutler, sd). Just like the MLR and DT method, there is one parameter that has to be determined first. For random forests, it is not feature variable based because it will choose the best automatically. It is also not the number of endpoints since these differ per generated tree. However, it is the number of variables randomly sampled as candidates at each split, also called mtry. Since the number of feature variables is six, the possible choices for the mtry are one to six. Utilizing bagging, the OOB error and test error determine the best mtry. Figure 12 gives an example of how it is visualised in R to make a decision.

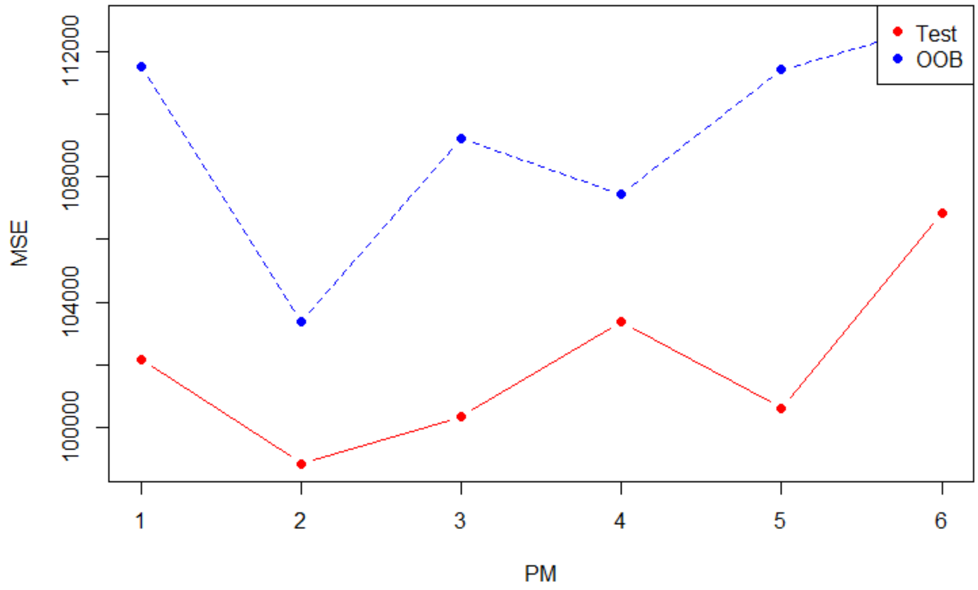


Figure 12: OOB error and test error to determine the best mtry.

Lower MSE values are better, so in this case, the best mtry is two. However, there may be no mtry that is minimum for both errors. In that case, the test error will lead, because the OOB error is more prone to overestimating (Janitza & Hornung, 2018). The determination of the best mtry enables the making of the final random forest model. Per model, the corresponding evaluation metrics indicate the quality. It is expected that the train and test MSE's are lower compared to the decision tree models. However, it is not possible to visualise the random forests because per project the average tree is likely to look different.

To provide more in-depth information, the tool contains code to determine the variable importance and quantiles. The variable importance has as input the random forest model which then determines what variable was used most in the random forest algorithm. Figure 13 shows the graph of the output of the VarImpPlot function. It shows that for that specific PM model, the variable Pieces is the most useful when predicting the number of hours. The x-axis stands for total decrease in node impurities.

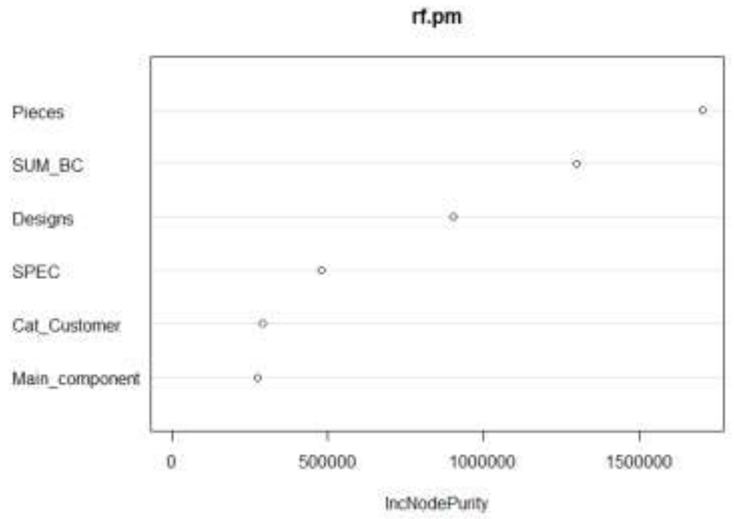


Figure 13: Output of the R function VarImpPlot.

The decision three method has an added analysis of the intervals of a certain node. This is also possible for RF, but in a slightly different way because there is not just one DT model within the RF model. But you can track all the values of the terminal nodes for every tree that is in the RF model. So, the intervals of the DT are based on the projects that walk their way to a single endpoint, while the intervals of the RF are based on all the estimates of the individual trees of which the average is the output of the RF model. So every project has their own interval and not every node. In this case, Quantile Regression Forest can help. QRF does not provide the intervals per terminal node, but an array containing all the endpoints throughout the randomly made individual trees. This array is treated the same as the decision tree array, where the array is sorted from high to low to show the quantiles of that array. The average is the estimate that follows from the random forest model. In the end, QRF is a method to look deeper than the average.

Figure 14 provides a summary of the steps taken to provide a model for a certain department. Again, the code handles the departments individually, but the same steps apply to every department.

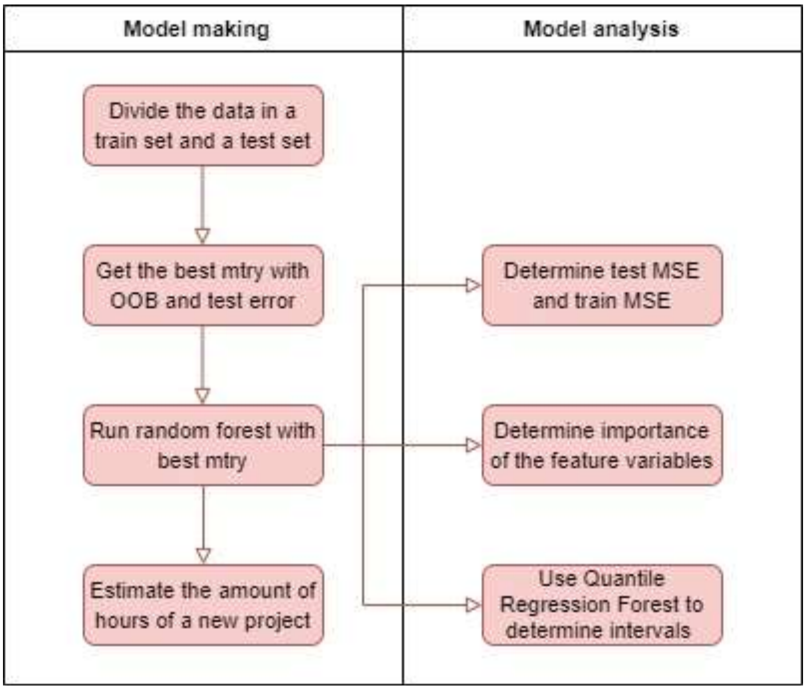

Model making
Model analysis
Divide the data in a train set and a test set
Get the best mtry with OOB and test error
Run random forest with best mtry
Estimate the amount of hours of a new project
Determine test MSE and train MSE
Determine importance of the feature variables
Use Quantile Regression Forest to determine intervals

Figure 14: Steps to make the random forest model.

4.3 Structure of the results

The three methods will provide main results and additional analysis opportunities per department. Section 4.1 addresses 45 projects which can serve as input. However, from 8 out of the 45 projects, the final number of hours are known. To not throw the realistic data away, it should be used in the dataset. However, it is also convenient to test all three models against the eight projects since the values are actuals. Therefore the results will have the following steps:

1. All three models are based on the dataset with the 45 projects estimated by VES throughout the bid phase. The methods how to model are stated in the last section;
2. The models and the estimates of VES are tested against the 8 finished projects with their realised values. Comparison is based on their MSE;
3. The realised values of the 8 finished projects replace the values of the estimates made during the bid phase in the dataset so it contains (45-8=) 37 estimated projects and 8 finished projects. This dataset is handled the same as in step 1 and will serve as the final models to estimate white-collar hours.

Step 1 and 2 are for validation, whereas step 3 contains the models for the tool for VES. The next chapter shows the MSE's per model for steps 1 and 2 and the MSE's plus the additional analysis for step 3. Table 7 summarises what data is shown in the next chapter for each department.

Table 7: Overview of the information per step for the next chapter.

Method	Step 1	Step 2	Step 3
MLR	Feature selection and MSE	MSE	Feature selection and MSE
DT	CV-depth selection and MSE	MSE	CV-depth selection, the model, MSE and intervals
RF	Mtry selection and MSE	MSE	Mtry selection, MSE and variable importance

5. Results

This chapter consists of four parts. The structure of section 5.1, 5.2, and 5.3 is the same as the steps described in the last section. They will show the characteristics and performance of the models. Section 5.4 provides a comparison between the models. The chapter ends with a note on validation. An important note is that the MSE is often a big number, therefore in the tables, the root of the MSE (RMSE) is shown to provide better intuition of the results. However, the MSE itself does not say a lot. It is dependent on the dataset. Table 8 functions to give more information about the departments to help understand the MSE value. The table includes the lowest value found in the dataset, the average of all projects, and the highest values found in the dataset. The main conclusion is that the max is often at least three times as high as the mean. This means that there are a couple of big projects that push the mean up, together with a set number of smaller projects that occur more often. The goal of the table is to interpret results between departments. For example, a root MSE of 500 for Engineering and MRP have entirely different meanings. To indicate this difference, the statistic RMSE divided by the mean is added.

Table 8: Characteristics of the dataset, based on the 45 estimates and 8 finished projects.

Department	Min (45)	Mean (45)	Max (45)	Min (8)	Mean (8)	Max (8)
PM	0	583	1701	420	558	987
Procurement	0	380	1757	161	322	494
Engineering	0	1133	6385	115	498	1408
Quality	20	892	4572	209	604	1088
Logistics	0	1096	5695	258	456	728
Planning	0	843	2823	276	735	1361
PP	25	330	868	35	203	429
MRP	24	260	812	50	148	267
WVB	24	433	1251	63	384	785

The min and max values stay quite the same, however, the means of the engineering and logistics department differ a lot. This is important to know when comparing the results of step 1 and step 3 since the new data will influence the models.

5.1 Models based on estimates

As mentioned in the last chapter, the three methods contain 45 estimates as the input dataset.

MLR

Table 9 shows the model selection and the RMSE. The model selection shows the selected features. The best performing model is the model for the department MRP, whereas the engineering department model performs the worst.

Table 9: Model characteristics MLR step 1.

Department	Model selection	RMSE (hours)	RMSE / Mean (%)
PM	SUM_BC	293	50,3
Procurement	Cat_Customer, Pieces, Design	203	53,4
Engineering	Cat_Customer, Design	975	86,1
Quality	All	352	39,5
Logistics	SUM_BC	481	43,9
Planning	SPEC, Pieces, Main_component, SUM_BC, Cat_Customer	335	39,7
PP	Cat_Customer,Main_component, SPEC, SUM_BC	177	53,6
MRP	All	82	31,5
WVB	SUM_BC	186	43,0

Decision tree

Table 10 shows the number of endpoints per department if pruning was necessary and the squared MSE. The best performing department is WVB and the worst-performing department is again engineering.

Table 10: Model characteristics for decision trees step 1.

Department	CV depth (endpoints)	Pruning	RMSE (hours)	RMSE /Mean (%)
PM	3	No	385	66,0
Procurement	4	Yes	347	91,3
Engineering	4	No	1412	124,6
Quality	5	No	672	75,3
Logistics	4	No	780	71,2
Planning	4	Yes	499	59,2
PP	4	No	252	76,4
MRP	3	No	160	61,5
WVB	4	No	235	54,3

Random forest

Table 11 shows the best mtry and both train and test error. This time the planning department performs the best, while the model for the engineering department is again the worst.

Table 11: Model characteristics for random forests step 1.

Department	Best mtry	Train RMSE	Test RMSE	Test RMSE / Mean (%)
PM	4	140	282	48,4
Procurement	4	144	267	70,3
Engineering	5	392	1727	152,4
Quality	2	254	428	48,0
Logistics	6	281	483	44,1
Planning	2	230	281	33,3
PP	1	122	242	73,3
MRP	2	61	123	47,3
WVB	2	116	223	51,5

5.2 Validation with the finished projects

In this step, the models made in step 1 predict the 8 projects of which the actual values exist. Table 12 shows the prediction quality of each method based on RMSE and the best method. Each dataset has its characteristics, such as linearity, and therefore differs. Intuitively, an RF model with a lot of trees should outperform a DT model, which is not the case in this scenario. This is likely caused by the importance of features variables. If only one is important, a random forest model will make a lot of trees that do not contain this variable. This harms the model, thus its (R)MSE will be higher.

Table 12: Comparison of the RMSEs based on the 8 projects validation set.

Department	Original estimates	MLR	DT	RF	Method with lowest root RMSE	Best RMSE / Mean (%)
PM	362	255	265	312	MLR	43,7
Procurement	201	164	185	186	MLR	43,2
Engineering	579	501	506	455	RF	40,2
Quality	452	428	318	406	DT	35,7
Logistics	458	465	368	391	DT	33,6
Planning	601	576	369	576	DT	43,8
PP	197	412	152	140	RF	42,4
MRP	104	407	70	79	DT	26,9
WVB	380	240	234	277	DT	54,0

The most important column to compare with is the original estimates column. For every department, at least one of the three methods outperforms. This means that the expert judgement of the individual departments is always a worse method to estimate then using data-driven methods. This is an important conclusion since it means that although the amount of data is small, still meaningful estimates with relation to their estimates result out of the three methods.

Most methods perform relatively equal. The main outliers are the MRP and PP departments, where the MLR method does a poor job estimating the number of hours. This likely implies that it is the only department that has no or almost no linearity in it. Another reason could be that those departments had less than 45 projects to work with, which made it harder to come up with a good model. However, this is not the case for WVB.

The last topic is the prediction power of the methods. The last column shows promising results, but the ratios are still quite big. When the RMSE is around 40% of the mean it means that it basically can't properly smaller projects. The dataset exists out of a lot of small or normal projects and a couple of large projects. The larger ones force the average to go up, therefore to properly predict smaller projects the ratio should be smaller. Chapter 6 continues on this.

5.3 Final models including the finished projects.

This section contains model selection results and MSE for the MLR model. The CV-depth selection, the tree, intervals and the MSE values for the decision tree model. Last, the mtry selection, MSE, and variable importance are part of the random forest model.

MLR

Table 13 shows the model selections and the RMSE.

Table 13: Model characteristics MLR step 3.

Department	Model selection	RMSE (hours)	RMSE / Mean (%)
PM	SUM_BC	300	51,5
Procurement	Designs	211	55,5
Engineering	Cat_Customer, Designs	991	87,5
Quality	SUM_BC, Designs	399	44,7
Logistics	SUM_BC, Pieces	454	41,4
Planning	All	380	45,1
PP	Designs	199	60,3
MRP	SUM_BC, Designs	92	35,4
WVB	SUM_BC	202	46,6

With the replacement of some of that data, the models change. For example, in step 1 all variables were included in the quality model, whereas now the sum of blue-collar and the number of designs serve as input. The best and worst performing departments stayed the same compared to step 1. However, the overall performance of all models has decreased a slight amount. The average RMSE of all departments is a little higher in step 3 than in step 1. This is not necessarily a bad thing as long as the difference is not too high. A new dataset means new models with new challenges, so improving the MSE is not granted.

Decision trees

Table 14 shows the number of endpoints per department t if pruning was necessary and the RMSE. Engineering has still the highest RMSE divided by the mean, whereas MRP now has the lowest. Figure 15 shows the trees, pruned if necessary for each department. Zooming in the first split of all the trees, the split SUM_BC < 8070 appears suspiciously often. This value does not appear in the dataset. The closest lower value is 7508 and the closest higher value is 8500. Apparently, the algorithm determines that this value is the best split to determine large and normal to small projects. However, this is not the case for logistics. Another remarkable result is the occurrence of the SUM_BC feature. Seven out of nine times it is used to make the first split, and four trees only use SUM_BC. This indicated that the algorithm only finds one useful feature. As addressed earlier, this might cause the DT model to perform better than the RF model. Finally, Table 15 shows as an example the intervals of the department PM that can be used as extra information. There is some overlap between the end of node 2 and the beginning of node 3. This indicates that the projects that have around 900 hours are more prone to ending up in the "wrong" node. Appendix C shows all the other intervals.

Table 14: Model characteristics DT step 3.

Department	CV depth (endpoints)	Pruning	RMSE (hours)	RMSE / Mean (%)
PM	2	Yes	396	67,9
Procurement	5	No	361	95,0
Engineering	4	No	1493	131,8
Quality	2	Yes	673	75,4
Logistics	4	No	753	68,7
Planning	4	Yes	469	55,6
PP	3	No	187	56,7
MRP	2	Yes	139	53,5
WVB	5	No	271	62,6

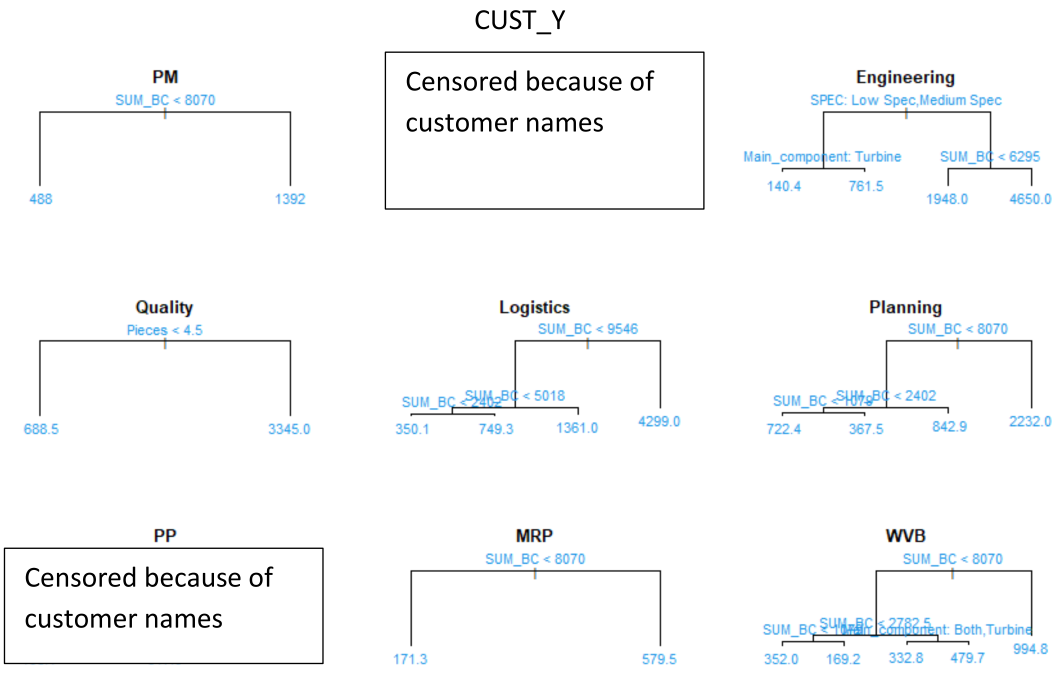


Figure 15: All tree models for step 3 visualised.

Table 15: Quantiles for the PM department. The lower the node number, the more right it node in the tree.

Quantile	Node 2	Node 3
0%	0	940
25%	350	1220,5
50%	453	1549
75%	568,5	1600
100%	1464	1616

Random forest

Table 16 shows the best mtry and both train and test error. Also in the last results engineering is the hardest to predict department. Planning this time has the lowest ratio. Generally, the train RMSE is lower than the test RMSE, but not for planning. This is nothing but a coincidence. Apparently, the planning model suits the test data very well.

Table 16: Model characteristics RF step 3.

Department	Best mtry	Train RMSE	Test RMSE	Test RMSE / Mean (%)
PM	2	170	318	54,5
Procurement	6	134	251	66,1
Engineering	1	888	1016	89,7
Quality	2	274	529	59,3
Logistics	4	322	427	39,0
Planning	1	350	319	37,8
PP	1	126	226	68,5
MRP	2	55	195	75,0
WVB	1	136	284	65,6

Figure 16 shows the variable importance.

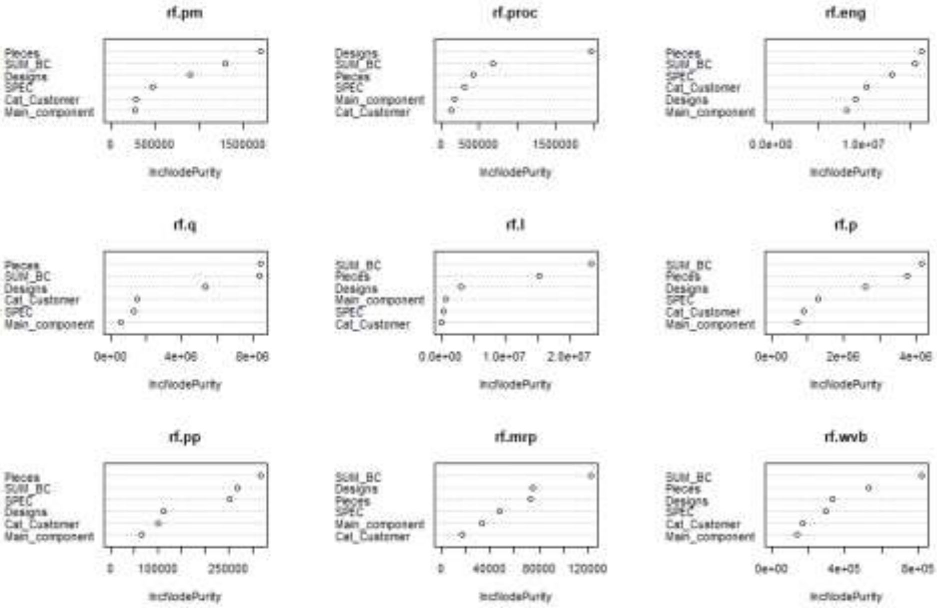


Figure 16: All VarImpPlot of RF step 3.

Again, the SUM_BC variable, the sum of all hours made in the workshop, influences almost all models. This does makes sense since a lot within VES is based on the workshop, their most important asset. For example, the more work to be done, the more planning. Four of the departments are planning departments, and are therefore influenced by sum of all blue-collar hours. Logistics and quality as well, the more work in the workshop, the more checks that need to be done and the more materials need to be moved through the company. The only model where SUM_BC is almost irrelevant is the procurement model. This does make sense since procurement is mostly influenced by the number of different parts that are necessary to finish the project. A bigger project with more pieces only means more pieces of the same part, which should not make procurement spend more hours on a project. Another important note is that pieces and SUM_BC are next to each other in eight out of the nine projects. The correlation between the two might indicate this link. The R^2 value of their correlation is 0,7368. The fact that they are not the same is the learning of the blue-collar workers in the workshop. If they have to work with new parts or regulations, it is expected that the first piece takes the longest time. If the project consists of more pieces, the blue-collar workers can apply their learnings to the next piece. This is of course assuming that the pieces are built after each other. Working in parallel, this is learning curve might be flattened because information needs to be transferred between workers to learn.

5.4 Comparison between the methods

Section 5.2 provides a first comparison of the methods based on the finished projects. This section bases the comparison on the RMSE results and features of Section 5.3

Comparison of RMSE

Table 17 shows the RMSE's of each method. MLR is six times the best method, RF two times and DT one time. The best performing model is the MLR method for the MRP department. The worst performing model is still the model for the engineering department. Either the methods chosen do not suit the data, or it is really hard to find good feature variables. These RMSE values are worst than the RMSE values in Section 5.2, but are these RMSE comparable? The RMSE's of the methods give an indication of the quality based on what is known. Even the RF test error is somewhat biased since its RMSE is still based on known information. And as stated before, the input is different which automatically changes the model. Thus, there is not much to learn making the comparison.

Table 17: Comparisons of the RMSE.

Department	MLR	DT	RF (test error)	Method with lowest RMSE	Best RMSE / Mean (%)
PM	300	396	318	MLR	51,5
Procurement	211	361	251	MLR	55,5
Engineering	991	1493	1016	MLR	87,5
Quality	399	673	529	MLR	44,7
Logistics	454	753	427	RF	39,0
Planning	380	469	319	RF	37,8
PP	199	187	226	DT	56,7
MRP	92	139	195	MLR	35,4
WVB	202	271	284	MLR	46,7

Comparison of the features

Each method has its unique feature variables. Table 18 provides an overview of the use of the features. The random forest column includes all the features until their first big gap in Figure 16. In case of a situation where there is no such gap, see PM and P, the best three variables are included.

Table 18: Features included in which model.

Department	MLR	DT	RF
PM	SUM_BC	SUM_BC	Pieces, SUM_BC, Designs
Procurement	Designs	SUM_BC, Pieces, Main_component, Cat_Customer	Designs
Engineering	Cat_Customer, Designs	SPEC, Main_component, SUM_BC	Pieces, SUM_BC, SPEC
Quality	SUM_BC, Designs	Pieces	Pieces, SUM_BC, Designs
Logistics	SUM_BC, Pieces	SUM_BC	SUM_BC, Pieces
Planning	All	SUM_BC	SUM_BC, Pieces
PP	Designs	SUM_BC, Cat_Customer	Pieces, SUM_BC, SPEC
MRP	SUM_BC, Designs	SUM_BC	SUM_BC
WVB	SUM_BC	SUM_BC, Main_component	SUM_BC

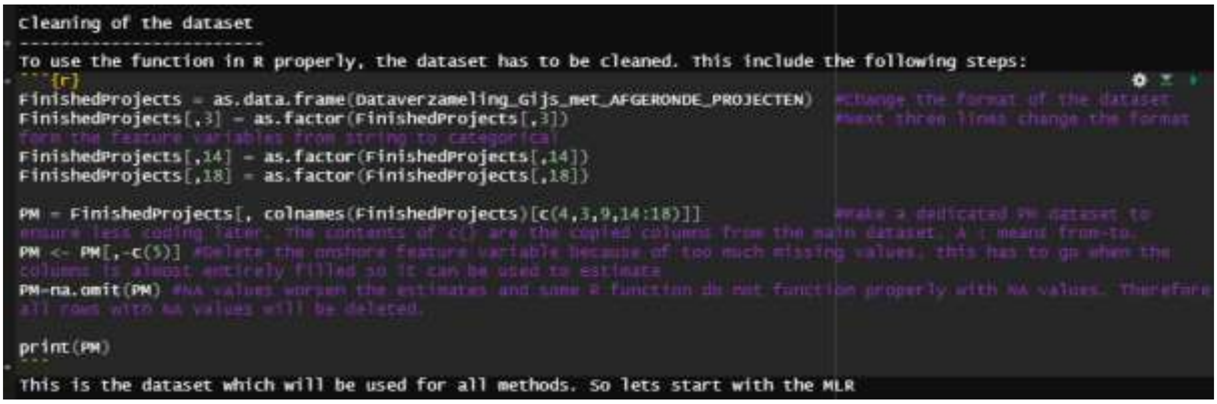
The most used feature variables are:

1. SUM_BC, 22 times;
2. Pieces, 10 times;
3. Designs, 9 times;
4. Cat_Customer, Main_component, and SPEC, all 4 times.

Based on this statistic, the total number of blue-collar hours is the most usable when estimating the number of white-collar hours. Pieces and designs are somewhat usable and the other three variables in a rare occasion. This does not mean you get better results when leaving one of the lesser-used ones out. The next question then is what happens when more features are added? The next chapter reflects on that.

5.5 The tool for VES

Getting information out of the tool starts with updating the excel sheet in Appendix B. Actual values should replace bid estimates as soon as possible. The influx of new data means that the models need to be updated. Therefore the tools follow the structure of Chapter 4. Every method has a dedicated R document that helps select the model, makes the model, and has multiple ways to evaluate the model. The employee that is going to work with the tool does need an understanding of R. For example, when they want to add an extra feature variable, some code has to be changed. This was agreed on by VES, and a technical sales manager will take over the model after the research is done. A R markdown document explains all the three methods, to make it easier to take over the code. R markdown is a way of exporting the code and comments to for example word. In this style you can explain the steps and the necessary inputs based on model selection which serve as the most important inputs. Figure X shows a screenshot of the code, while Appendix X shows the same code exported to word. This was agreed on by VES, and a technical sales manager will take over the model after the research is done.



~~~
Cleaning of the dataset
---------------------------
To use the function in R properly, the dataset has to be cleaned. This include the following steps:
```{r}
FinishedProjects = as.data.frame(Dataverzameling_Gijs_met_AFGERONDE_PROJECTEN)   #Change the format of the dataset
FinishedProjects[,3] = as.factor(FinishedProjects[,3])                           #Next three lines change the format
form the feature variables from string to categorical
FinishedProjects[,14] = as.factor(FinishedProjects[,14])
FinishedProjects[,18] = as.factor(FinishedProjects[,18])

PM = FinishedProjects[, colnames(FinishedProjects)[c(4,3,9,14:18)]]              #Make a dedicated PM dataset to
ensure less coding later. The contents of c() are the copied columns from the main dataset. A : means from-to.
PM <- PM[,-c(5)] #Delete the onshore feature variable because of too much missing values. this has to go when the
columns is almost entirely filled so it can be used to estimate
PM=na.omit(PM) #NA values worsen the estimates and some R function do not function properly with NA values. Therefore
all rows with NA values will be deleted.

print(PM)
```
This is the dataset which will be used for all methods. So lets start with the MLR
~~~

Figure 17: Screenshot from the tool.

5.6 Validation and sensitivity analysis

Normally validation and sensitivity make up a big part of every research. However this case is a little uncommon. The current dataset and methods provide models and results based on the analysis of the models but as soon as the dataset is changed, the model selection (e.g. best mtry, cv-depth, feature selection), the models, and results change as well. Next to this, some validation is included in the methods. Based on the book that served as a guideline (James, Witten, Hastie, & Tibshirani, 2017), analysis after the model building is rarely done. Both factors make a sensitivity analysis hard to do. After some research, the most common sensitivity analysis in statistical learning is determining the size of the dataset. This the data are scarce, this makes no sense to do.

The size of the dataset also influences the possibilities of validation. In the model building process, choices have to be made that influence the validation. For example backwards or forwards selection when determining the model. However, small data makes this choice redundant since both end up as the same. Appendix D shows an example. All three models are already getting a lot out of the data, so the change in validation changes cause no impact on the final model.

6. Conclusion, discussion and recommendations

First, the conclusion provides a broad perspective on the results. Second, the discussion addresses the features and the projects. The chapter ends with the recommendations for VES since further research on this specific topic is hard.

Conclusion

Chapter 5 explains the stand-alone results of the methods. The models determine the most impactful conclusion; how well do the methods do? The answer to that question is two-fold. Firstly, the models made in step one, based on the estimates made throughout the sales process, have generally a lower RSME then the estimates the individual departments made for the finished projects. This means that using all old estimates is a better way than the currently often used filling in an excel sheet. Secondly, the individual RMSE values are not enough to give a reliable estimate. The main cause of this is the small amount of data. Most statistical learning methods are suitable for big data, which makes sense, but it is not always the case within a company. So based on the RMSE we can conclude that the three methods outperform expert judgement, but that the RMSE has a lot of room to improve.

Therefore the use of the methods will change when VES keeps updating the data. Right now the predicting power needs to be judged with a grain of salt, but as time passed and more projects enter the dataset, the white-collar estimates will become more like reality and thusly better. The first way to collect more data is to add new projects and to fill in data of finished projects. A second extension could be the addition of more feature variables, like a complete onshore variable. The fact that expert judgement scored worst does not mean it always gives the worst estimate. It might happen that the RMSE of the three methods is low, but that expert judgement is still the best method. The ability for the models to learn is likely still less than the intuition some of the employees have.

Maybe with more data, it can even happen that one method ends up being the clear favourite. Right now, the results show no clear winner between the methods. Even the worst method based on RMSEs, the decision tree, has a use. For some departments, the decision tree performs slightly worse than MLR or RF. In that case, the interpretability of the decision tree methods might cause VES to favour it instead of the method with the best RMSE.

Discussion

The discussion consists of three main parts. The approach, the dataset and the features. The methods are not addressed since Section 5.6 discusses them. Small data just limits the number of possible methods that change the predictive power.

The approach

The main discussion point of this approach is the use for the project management department. The first problem was budget overruns, which this approach does not directly tackle. It indicates the way individual departments should work, but not their interaction. Ideally, the time a department spends on a project is only based on the tasks of that department and not influenced by how departments collaborate. In real life, every hour of work is written on a project. Every hour of bad communication, human errors, or other delays end up in the dataset. These factors are, hopefully not, dependant on the features the model use and therefore create a certain noise. The impact of this is not known and it could be interesting to research.

The dataset

One of the main disadvantages of the current dataset is that every project has the same influence on the model. An extension of the dataset could be a certain weight to indicate the quality of a finished project with overruns. A part of the increase of the final hours compared to the bid might be caused by human interaction within the company or by the collaboration with the customer. For example, a project manager gets 400 hours for a project, but halfway through the project, the contact person of the company leaves the company without proper handover. Then someone else has to fill the empty spot but misses information, which ultimately decreases the quality of the collaboration. This might cause a budget overrun for the project manager that could not be prevented. Right now projects like this have the same impact on the model then perfectly executed projects. It would be interesting to see what happens when this is included, but the main obstacle would be how to rate the project. An argument against this idea might be that these influences are part of the real world, and the weights are a bad idea. However, the estimations serve to make a bid on a project and if all the possible misfortunes are incorporated in the model it likely becomes too expensive for the possible customer.

Next, one of the main topics is small data and the difficulties working with it. One of the ways to enlarge the dataset would be with data augmentation, using the data you have to make more data. This method has not been used because of two reasons. The main reason is that the current analysis can be redundant tomorrow, because of the addition of new data. Therefore the determination of the best augmentation process that suits the bias of the dataset would mean an addition to the tool. When dataset get larger, it would even be possible that augmentation is not necessary anymore and therefore, together with the time set for the research, augmentation is not included. However, it is a possibility for further research to see if there is anymore to get out of the data.

The features

The features used in the analysis are very basic. This has an advantage and a disadvantage. The advantage is that early in the sales process enough information is available to estimate the number of hours. The disadvantage is that it is hard to find good relations between the features and the data and therefore impact the quality of the model negatively. More specific features, open up possibilities to change the quality of the models because of the inclusion of different correlations.

Recommendations

The first recommendation is to keep using the model as a conversation starter during the sales process. The expert judgement method is proven, based on RMSE, to perform in any case not better than the three methods combined. Ideally, the expert judgement methods are improved with the findings from the methods, to ensure a more realistically sold project. This starts with taking a look at the estimation methods used by the departments, often a small excel tool. The ability to define which actions need to be in a certain project is something VES is most likely doing well. However, every action has a set number of hours which are rarely adjusted after a project is done. This could be a second way of learning, next to the learning the three models will do.

The second recommendation is an obvious one; improve the dataset. 45 projects is a shallow number to work with and even the majority is not based on the performance of the company but based on expert judgement. Improving the dataset can be done with more entries, but as the discussion stated also with weights. The next recommendation is to properly keep track of information per project so the number of features can extend. It was hard to get information that should be easy to get by. VES also knows this already, but their ERP system and document control can use some work to make future data analysis more straightforward to do. The last recommendation is to reconsider the methods and model selection when a total of 100 projects are in the datasheet. It does not have to be all finished projects, but it might cause some decision-making to change because of the available models and the likelihood of model selection to become more difficult.

References

Araya, F., Faust, K., Khwaja, N., O'Brien, W. J., Liang, X., & Bur, M. K. (2020). Exploring a quantitative and qualitative mixed approach for estimating preliminary engineering efforts of bridge replacement projects. *Transportation Research*, 13-22. doi:10.1177/0361198120917677

Awada, M., Srour, F. J., & Srour, I. M. (2021). Data-driven machine learning approach to integrate field submittals in project scheduling. *Journal of Management in Engineering*. doi:10.1061/(ASCE)ME.1943-5479.0000873

Breidman, L. (2001). Random forests. *Machine Learning*, 5-32.

Breiman, L., & Cutler, A. (sd). *Random Forests*. Opgehaald van stat.berkeley: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#ooberr

Elmousalami, H. H. (2021). Comparison of artificial intelligence techniques for project conceptual cost prediction: A case study and comparative analysis. *IEEE Transactions on engineering management*, 183-196. doi:10.1109/TEM.2020.2972078

Fandopa, R., Alisjahbana, S. W., & Ma'soem, D. M. (2020). The analysis of factors that affect the cost of the soil work of the road project. *International journal of advanced science and technology*, 1258-1267.

Gonfalonieri, A. (2019, June). *5 ways to deal with the lack of data in machine learning*. Opgehaald van KDnuggets: https://www.kdnuggets.com/2019/06/5-ways-lack-data-machine-learning.html

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning.* Stanford, California: Springer.

Heerkens, H., & van Winden, A. (2012). *Geen probleem.* Vwc.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning with Applications in R.* Standford, California: Springer-Verlag New York Inc.

Janitza, S., & Hornung, R. (2018). On the overestimation of random forest's out-of-bag error.

Meinshausen, N. (2006). Quantile Regression Forests. *Journal of Machine Learning Research*, 983–999.

Öztürk, A., Kayaligil, S., & Özdemirel, N. E. (2006). Manufacturing lead time estimation using data mining. *European journal of operational research*, 683-700. doi:10.1016/j.ejor.2005.03.015

Pahno, S., Yang, J. J., & Kim, S. S. (2021). Use of machine learning algorithms to predict subgrade resilient modulus. *Infrastructures*. doi:https://doi.org/10.3390/infrastructures6060078

Persad, K. R., O'Connor, J. T., & Varghese, K. (1995). Forecasting engineering manpower requirements for highway preconstruction activities. *Journal of Management in Engineering*, 41-47.

Rocca, J. (2019, April 23). *Ensemble methods: bagging, boosting and stacking*. Opgehaald van Towards data sciende: https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

Rothmann, D. (2019, July). *7 tips for dealing with small data*. Opgehaald van KDnugget: https://www.kdnuggets.com/2019/07/7-tips-dealing-small-data.html

Shepperd, M., & Cartwright, M. (2001). Predicting with sparse data. *IEEE Transactions on software engineering*, 987-998.

Statistical Learning Theory. (sd). Opgehaald van deepai.org: https://deepai.org/machine-learning-glossary-and-terms/statistical-learning-theory

Tokuç, A. (2021, January 14). *Splitting a Dataset into Train and Test Sets*. Opgehaald van baeldung.com: https://www.baeldung.com/cs/train-test-datasets-ratio

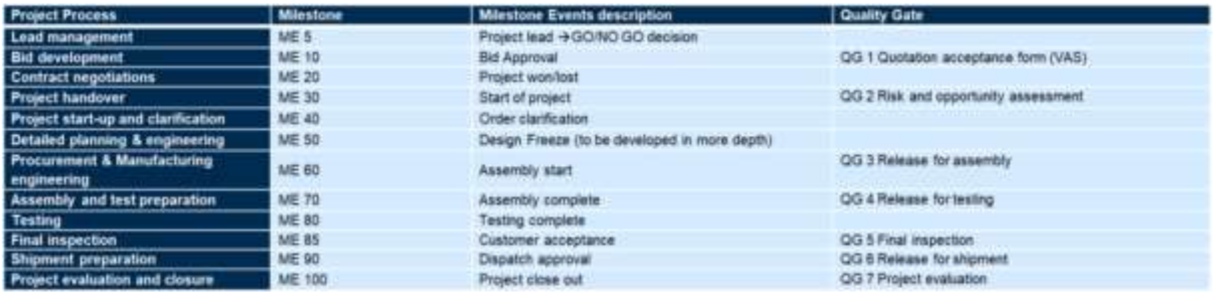
Varshini, A. P., Kumari, K., Janani, D., & Soundariya, S. (2021). Comparative analysis of machine learning and deep learning algorithms for software effort estimation. *Journal of Physics: Conference Series*. doi:10.1088/1742-6596/1767/1/012019

Varshini, P. A., Kumari, A. K., & Varadarajan, V. (2021). Estimating software development efforts using a random forest-based stacked ensemble approach. *Electronics*. doi:https://doi.org/10.3390/

Zhang, Z. (2016, June 26). *Boosting algorithms explained*. Opgehaald van Towards data science: https://towardsdatascience.com/boosting-algorithms-explained-d38f56ef3f30

Appendices

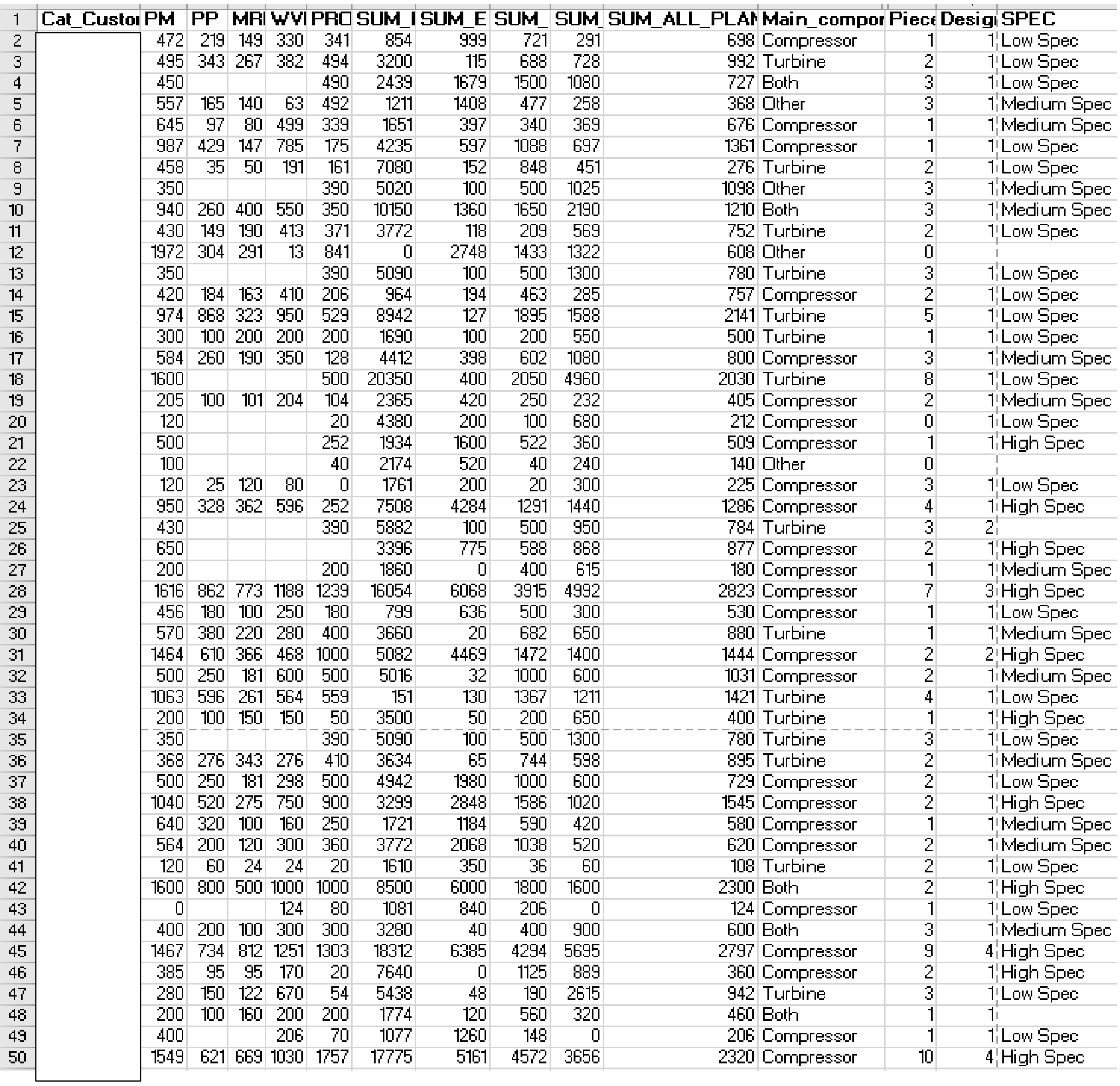
Appendix A: Milestone format of VES



Project Process	Milestone	Milestone Events description	Quality Gate
Lead management	ME 5	Project lead →GO/NO GO decision	
Bid development	ME 10	Bid Approval	QG 1 Quotation acceptance form (VAS)
Contract negotiations	ME 20	Project won/lost	
Project handover	ME 30	Start of project	QG 2 Risk and opportunity assessment
Project start-up and clarification	ME 40	Order clarification	
Detailed planning & engineering	ME 50	Design Freeze (to be developed in more depth)	
Procurement & Manufacturing engineering	ME 60	Assembly start	QG 3 Release for assembly
Assembly and test preparation	ME 70	Assembly complete	QG 4 Release for testing
Testing	ME 80	Testing complete	
Final inspection	ME 85	Customer acceptance	QG 5 Final inspection
Shipment preparation	ME 90	Dispatch approval	QG 6 Release for shipment
Project evaluation and closure	ME 100	Project close out	QG 7 Project evaluation

Figure 18: Milestones in a packaging project

Appendix B: The dataset



1	Cat_Custo	PM	PP	MRI	WVI	PRO	SUM_I	SUM_E	SUM_	SUM_	SUM_ALL_PLAN	Main_compor	Piece	Desig	SPEC
2		472	219	149	330	341	854	999	721	291	698	Compressor	1	1	Low Spec
3		495	343	267	382	494	3200	115	688	728	992	Turbine	2	1	Low Spec
4		450				490	2439	1679	1500	1080	727	Both	3	1	Low Spec
5		557	165	140	63	492	1211	1408	477	258	368	Other	3	1	Medium Spec
6		645	97	80	499	339	1651	397	340	369	676	Compressor	1	1	Medium Spec
7		987	429	147	785	175	4235	597	1088	697	1361	Compressor	1	1	Low Spec
8		458	35	50	191	161	7080	152	848	451	276	Turbine	2	1	Low Spec
9		350				390	5020	100	500	1025	1098	Other	3	1	Medium Spec
10		940	260	400	550	350	10150	1360	1650	2190	1210	Both	3	1	Medium Spec
11		430	149	190	413	371	3772	118	209	569	752	Turbine	2	1	Low Spec
12		1972	304	291	13	841	0	2748	1433	1322	608	Other	0		
13		350				390	5090	100	500	1300	780	Turbine	3	1	Low Spec
14		420	184	163	410	206	964	194	463	285	757	Compressor	2	1	Low Spec
15		974	868	323	950	529	8942	127	1895	1588	2141	Turbine	5	1	Low Spec
16		300	100	200	200	200	1690	100	200	550	500	Turbine	1	1	Low Spec
17		584	260	190	350	128	4412	398	602	1080	800	Compressor	3	1	Medium Spec
18		1600				500	20350	400	2050	4960	2030	Turbine	8	1	Low Spec
19		205	100	101	204	104	2365	420	250	232	405	Compressor	2	1	Medium Spec
20		120				20	4380	200	100	680	212	Compressor	0	1	Low Spec
21		500				252	1934	1600	522	360	509	Compressor	1	1	High Spec
22		100				40	2174	520	40	240	140	Other	0		
23		120	25	120	80	0	1761	200	20	300	225	Compressor	3	1	Low Spec
24		950	328	362	596	252	7508	4284	1291	1440	1286	Compressor	4	1	High Spec
25		430				390	5882	100	500	950	784	Turbine	3	2	
26		650					3396	775	588	868	877	Compressor	2	1	High Spec
27		200				200	1860	0	400	615	180	Compressor	1	1	Medium Spec
28		1616	862	773	1188	1239	16054	6068	3915	4992	2823	Compressor	7	3	High Spec
29		456	180	100	250	180	799	636	500	300	530	Compressor	1	1	Low Spec
30		570	380	220	280	400	3660	20	682	650	880	Turbine	1	1	Medium Spec
31		1464	610	366	468	1000	5082	4469	1472	1400	1444	Compressor	2	2	High Spec
32		500	250	181	600	500	5016	32	1000	600	1031	Compressor	2	1	Medium Spec
33		1063	596	261	564	559	151	130	1367	1211	1421	Turbine	4	1	Low Spec
34		200	100	150	150	50	3500	50	200	650	400	Turbine	1	1	High Spec
35		350				390	5090	100	500	1300	780	Turbine	3	1	Low Spec
36		368	276	343	276	410	3634	65	744	598	895	Turbine	2	1	Medium Spec
37		500	250	181	298	500	4942	1980	1000	600	729	Compressor	2	1	Low Spec
38		1040	520	275	750	900	3299	2848	1586	1020	1545	Compressor	2	1	High Spec
39		640	320	100	160	250	1721	1184	590	420	580	Compressor	1	1	Medium Spec
40		564	200	120	300	360	3772	2068	1038	520	620	Compressor	2	1	Medium Spec
41		120	60	24	24	20	1610	350	36	60	108	Turbine	2	1	Low Spec
42		1600	800	500	1000	1000	8500	6000	1800	1600	2300	Both	2	1	High Spec
43		0			124	80	1081	840	206	0	124	Compressor	1	1	Low Spec
44		400	200	100	300	300	3280	40	400	900	600	Both	3	1	Medium Spec
45		1467	734	812	1251	1303	18312	6385	4294	5695	2797	Compressor	9	4	High Spec
46		385	95	95	170	20	7640	0	1125	889	360	Compressor	2	1	High Spec
47		280	150	122	670	54	5438	48	190	2615	942	Turbine	3	1	Low Spec
48		200	100	160	200	200	1774	120	560	320	460	Both	1	1	
49		400			206	70	1077	1260	148	0	206	Compressor	1	1	Low Spec
50		1549	621	669	1030	1757	17775	5161	4572	3656	2320	Compressor	10	4	High Spec

Appendix C: Intervals for DT for step 3

Proc	3	6	7	8	9
0%	20	0	20	252	350
25%	80	20	318,5	330	514,5
50%	200	104	390	500	1000
75%	252	128	490,5	700	1271
100%	400	206	559	1000	1757

Eng	3	4	6	7
0%	20	0	50	0
25%	100	200	775	4503,25
50%	115	597	1600	5580,5
75%	130	1260	2848	6051
100%	400	2068	4469	6385

Q	2	3
0%	20	1895
25%	317,5	2050
50%	555	3915
75%	1009,5	4294
100%	1800	4572

L	4	5	6	7
0%	0	520	451	2190
25%	245	600	1093,75	3656
50%	300	680	1350	4960
75%	394,5	884	1551	4992
100%	1211	1080	2615	5695

P	4	5	6	7
0%	206	108	212	1210
25%	530	191,25	673,5	2085,5
50%	698	386,5	800	2300
75%	757	506,75	1011,5	2558,5
100%	1421	676	1545	2823

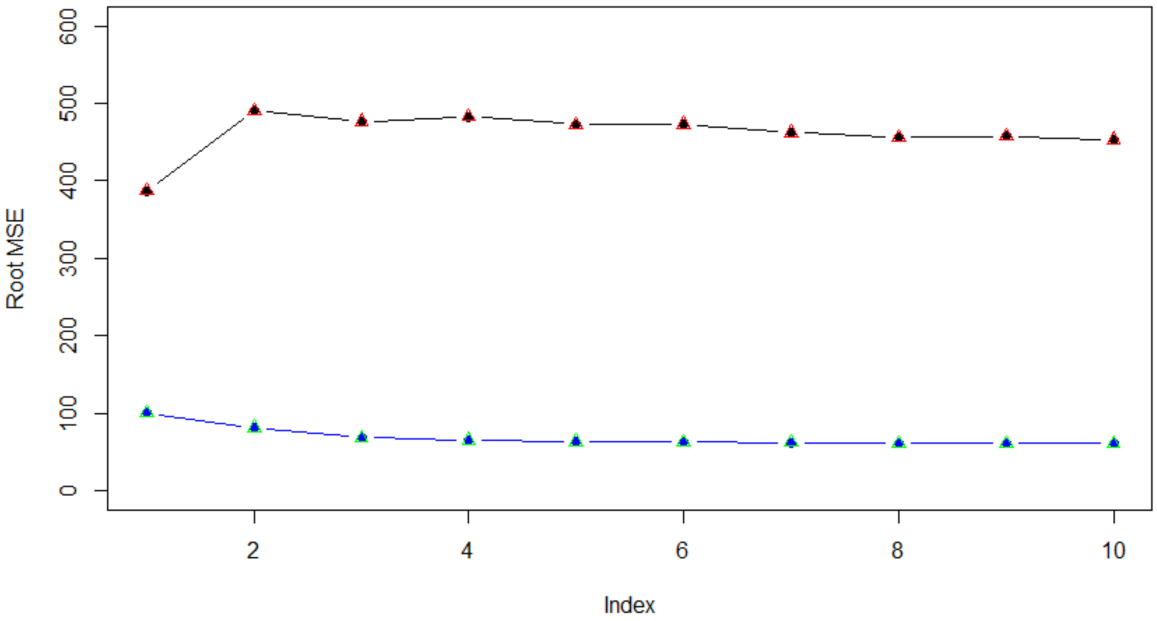
PP	3	4	5
0%	25	180	260
25%	98,5	200	649,25
50%	150	250	767
75%	268	328	846,5
100%	596	610	868

MRP	2	3
0%	24	323
25%	100,75	425
50%	149,5	584,5
75%	205	747
100%	366	812

WVB	4	5	7	8	9
0%	206	24	150	170	550
25%	250	75,75	254,75	300	962,5
50%	330	142	290	468	1015
75%	410	201	389,75	600	1148,5
100%	564	499	670	785	1251

Appendix D

The black and blue lines and points resemble a forwards approach. The red and green triangles are the backwards approach and perfectly overlap the blue and black ones.



Appendix E

This is a part out of the R markdown document that serves as a tool to help the sales department estimate like done in this thesis.

Cleaning of the dataset

To use the function in R properly, the dataset has to be cleaned. This include the following steps:

```
FinishedProjects = as.data.frame(Dataverzameling_Gijs_met_AFGERONDE_PROJECT
EN)  #Change the format of the dataset
```

```
FinishedProjects[,3] = as.factor(FinishedProjects[,3])
#Next three lines change the format form the feature variables from string
to categorical
FinishedProjects[,14] = as.factor(FinishedProjects[,14])
FinishedProjects[,18] = as.factor(FinishedProjects[,18])

PM = FinishedProjects[, colnames(FinishedProjects)[c(4,3,9,14:18)]]
#Make a dedicated PM dataset to ensure less coding later. The contents of c
() are the copied columns from the main dataset. A : means from-to.
PM <- PM[,-c(5)] #Delete the onshore feature variable because of too much m
issing values, this has to go when the columns is almost entirely filled so
it can be used to estimate
PM=na.omit(PM) #NA values worsen the estimates and some R function do not f
unction properly with NA values. Therefore all rows with NA Values will be
deleted.

print(PM)
```

This is the dataset which will be used for all methods. So lets start with the MLR