Vaasan yliopisto
UNIVERSITY OF VAASA

Patrick Koivunen

Project Cost Plan Forecasting in Marine Services Industry

The School of Technology and Innovations
Master's Programme in Industrial Management

Vaasa 2022

UNIVERSITY OF VAASA
The School of Technology and Innovations

Author:	Patrick Koivunen
Name of Thesis:	Project Cost Plan Forecasting in Marine Services Industry
Degree:	Master of Science in Economics and Business Administration, Industrial Management
Programme:	Master's Programme in Industrial Management
Supervisor:	Petri Helo
Year of graduation:	2022 **Pages:** 82

ABSTRACT:

This thesis investigates difficulties of project cost planning for a global organization in marine services industry. The research was conducted with co-operation with the organization, and required datasets were gathered from organization’s data warehouse.

Forecasting of a cost plan is a difficult task where the intention is to predict the unknown risks and wrong estimations in cost plan causes difficulties and uncertainty when risks arise. The main objective of this thesis is to find common fundamental characteristics for positive and negative cost overruns affecting the cost plans and investigate how to better identify risky project characteristics for more accurate cost planning.

Previous research with the exact same objectives in the same industry area are in few, if any. However, similar previous research states forecast and planning difficulties. Methods of this study was a combination of both qualitative and quantitative research methods. The data was gathered from the enterprise data warehouse and the selected dataset consisted of around 2500 projects started in marine services industry from beginning of 2018 to end of 2020. For better investigation of this dataset, a data-driven multi-dimensional data model was created where a regressor model was created to forecast new cost plan values.

In the research it was highlighted that the risky project characteristics can be identified to certain segments, customer countries, vessel types, product types and project managers, and there are some correlations between these different risky characteristics. These findings are impactful as it is crucial for business to identify possible risky types of projects where the negative cost overruns are highly likely.

KEY WORDS: Marine, Services, Project, Cost, Plan, Overrun, Forecast

Table of Contents

Pictures

Figures

Tables

Abbreviations

ACWP	**Actual Cost of Work Performed**
BCWS	**Cost Planed Cost of Work Scheduled**
BCWP	**Cost Planed Cost of Work Performed**
EDW	**Enterprise Data Warehouse**
EEDI	**Energy Efficiency Design Index**
ETL	**Extract, Transform and Load**
ERP	**Enterprise Resource Planning**
GHG	**Greenhouse Gas**
HFO	**Heavy Fuel Oil**
LNG	**Liquefied Natural Gas**
IMO	**International Maritime Organization**
P1	**Initial Project Cost Plan**
P0	**Actual Project Cost Plan**
VLSFO	**Very Low Sulphur Fuel Oil**
SOLAS	**Safety of Life at Sea**
SQL	**Structured Query Language**

1 Introduction

Awareness of global warming has become part of our everyday life. This global change in the world is now forcing the humankind to innovate new solutions in a transition into emission free world while continuing with economic feasibility. In maritime industry, the potential average lifetime of a modern vessel is about 25 - 30 years, therefore the ship owners have a certain business need to think about economically feasible solutions which will last long for future needs in regards of efficiency and emissions. Luckily, majority of the vessels can be recycled when the lifetime of the vessel comes to an end. However, in most of the cases, a more economical and environmentally friendly solution is to retrofit the vessel's equipment by service projects to meet lower emission standards instead of scrapping down the vessel.

In January 2018, there were approximately 94 171 commercial seagoing vessels in the world (Review of Maritime Transport, 2018). By January 2022 there was a significant rise to approximately 99 800 commercial seagoing vessels (Review of Maritime Transport 2021). The shipping industry is growing, by the ever-growing global economic world but also as awareness of environmentally friendly solutions is growing.

The International Maritime Organization (IMO) has administrated conventions to ensure maritime safety and limit environmental effects. According to the SOLAS, International Convention for the Safety of Life at Sea (1973), administrated by the IMO every sea-going vessel needs to be drydocked in every five years. Drydocking is a complex, time consuming and expensive operation which requires a lot of planning from the ship owner and personnel. Some vessels have a special privilege for extended dry-docking and therefore have their dry-docking period from 5 to 7,5 years. In these drydocking periods, service projects are conducted to check and repair the vessel and upgrade equipment of the vessel.

The MARPOL, International Convention for the Prevention of Pollution from Ships (1974), is one of the most important marine environmental conventions by IMO. MARPOL convention strives to reduce to minimum the pollution of the oceans and seas. These conventions are also one of the reasons for ship owners to reduce greenhouse gas emissions with service projects and strive for completely emission free solutions.

Majority of marine services projects are upgrade or performance work done on an existing product or a system for an existing sea-going vessel by the organization this research was conducted for. However, there can be also totally new products or systems installed to the vessel. Depending on the scope of the work to be done on the vessel, these marine services projects are completed while the vessel is dry-docked, in a shipyard or while vessel is sailing.

The specific organization in marine services industry this thesis is conducted for, has two different cost plans set for the marine service project. The initial cost plan estimate of the project is set by the sales team when offering to the customer and the final cost plan set by the project team as the project was executed. Between these two, the initial and final cost plans, for the same project, are overruns. These cost overruns can be either positive or negative cost overruns. The purpose of this thesis is to find root causes for these overruns based on project data available.

The project cost plan set for a project is basically a fund estimated during project planning phase based on forecast on how much the project is expected to cost at completion. This forecast is based on calculation estimates. During the execution of the project this fund is followed but positive or negative deviations to the initial plan are likely to happen.

Due to lack of information, unknown risks, and unique nature of individual projects, it is usually a very difficult task to estimate or forecast a correct initial project cost plan estimate accurately. However, the initial cost plan will set the framework for the project manager, project engineers and rest of the project team to work on and is used to estimate the availability of needed resources.

1.1 The Objective of the Study and Research Questions

The research gap indicates that there are differences between initial cost plans and final cost plans in projects. These differences might be caused by different characteristics related to the project. This research investigates those characteristics.

The importance of this thesis could be seen as an organizational advantage to learn from previous projects and to forecast more correctly the initial cost plans and to also identify the risky projects which all could lead to better profitability.

In this research study a research questions are used to find out answer for the research problem. The research problem of the research is:

What is causing differences between initial and completed project cost plans?

The research questions used in the thesis to answer the research problem are such as:

Which type of project the marine service projects are?

What is the accuracy of the initial project cost plan estimates?

How to identify risky project characteristics?

1.2 Research Data and Research Methods

This research paper is an empirical study where the quantitative and qualitative data is collected from organization’s internal enterprise data warehouse. The quantitative research data consists of project data used belonging to the marine services industry and is used as a baseload for the analysis. The qualitative research data is collected from already executed projects where all data is completed.

The data of this thesis was collected from the organization's Enterprise Data Warehouse (EDW) and mostly presented through Microsoft PowerBi visualizations. The data is gathered from data tables in the EDW fact tables and corresponding dimensional tables. These data tables were joined with each other with SQL commands and one-to-one, one-to-many and many-to-one relationship connections by using data keys of the data tables.

1.3 Scope and limitations of the Study

The scope of this research is limited to the selected organization's marine service projects only. Marine services projects are of unique nature, which may limit and lead to some results that are not unambiguous.

The data scope selected for the research is limited to certain product data used in the projects, to consider all possible products or product upgrades supplied by the business unit in the organization. Therefore, the study does not reflect the whole organization, but one of its leading business units taking care of service projects.

The service projects selected into the scope are started from beginning of 2018 to end of 2020 to limit the number of projects to around 2500. The number of projects in this time period is a bit higher than historical number of projects as there is a globally growing need of service projects in vessels. This scope was selected in order to get an enough large dataset with as much as possible up to date data. This number of projects also gives a good variety of results as the nature of the business may be sensitive to global economic trends.

Years from 2018-2019 are representing projects before the COVID-19 pandemic and projects started in 2020 are to some extend affected by the pandemic. On the financial side, COVID-19 pandemic affected project personnel costs in case of quarantine costs and has pushed some deadlines into the future. As said, during 2018-2020, the average number of projects per year is a bit higher than earlier years. All of these provide the research with a diverse dataset of project cost plans.

1.4 Structure of the Study

This research consists of five main chapters. These chapters investigate the topic on different angles. After introduction the focus will be on the latest related literature review on the topic. The main themes in literature review are nature of projects, planning of projects, estimation of projects and business intelligence.

After the literature review the methods of the research are documented and explained which kind of methods are used. This chapter enables the replication of the research as much as it is possible. In the results section the detailed observations from dataset are documented. The last chapter returns to conclude the main themes of the research and reflections on the future research possibilities.

2 Literature Review

In this chapter we get more in-depth with the previously written literature related to the thesis themes. The data of literature contains studies, journals, and latest research. The main themes of the literature research include topics such as multi-stakeholder projects, service projects, marine projects, planning of projects, project risks, forecasting project costs and aspects of utilization of business intelligence.

2.1 Projects

Chich, Zwikael, Restubog (2019) explains that a project is often considered as a temporary organization structure of different internal and external resources and stakeholders aimed at achieving specific targeted goals. Therefore, one could say that each project remains always unique as there are always some type of variation in project specifics. Today, organizations are increasingly recognizing projects are value creation mechanisms for the organization. However, the value of the project can be subjectively perceived, and the value should be created together by the different stakeholders of the project. In regards of these multi-stakeholder projects, trust between the companies is of course an important factor in enabling of value creation.

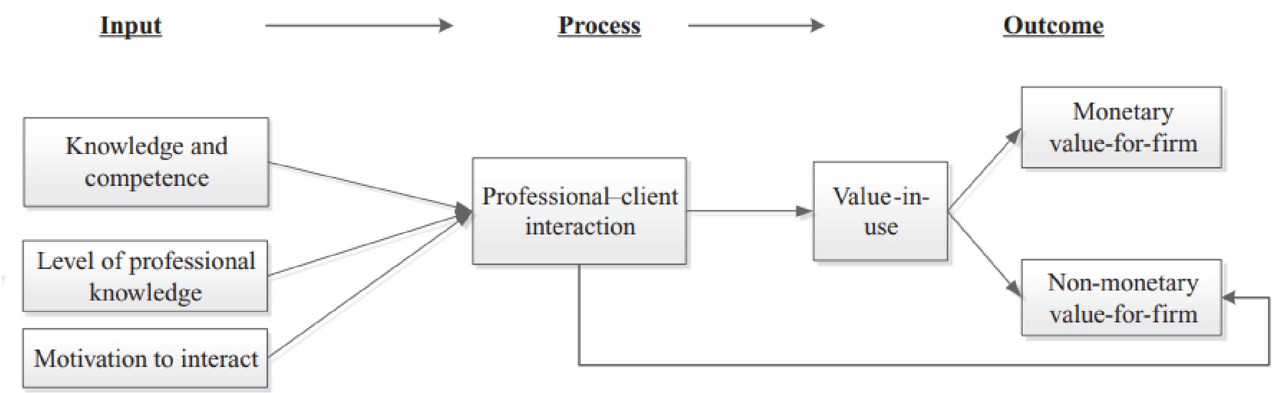


Figure 1 Value co-creation mechanism in professional service projects (Chich, Zwikael, Restubog (2019).

Artto, Kujala, Dietrich & Martinsuo (2008) confirms that the project goals as well as project management approaches of different stakeholders of the project may be very different. As especially multi-stakeholder projects with internal and external stakeholders, are inherently complex, and the project needs to consider the different interests of different stakeholders when setting a goal for the project. Project cannot therefore directly adopt into one suit and one strict goal communicated by the top management from one single organization. Therefore, the project needs really to position itself clearly according to the project's environment and project goals need to be always matched with the current situation and context.

2.1.1 Service Projects

As Chich et al (2019) explains, a professional service project is in a critical role in enhancing organizational performance. This is because in a service project, the project manager needs to work collaboratively with the customer to firstly define the problem, produce and deliver the service solution that will solve customer's problem. As the projects are of unique nature, the project value and the associated service package is often not specified well in advance but rather the value of the project is shaped by the inputs from different professionals and customer.

Therefore, as Chich et al (2019) continues, the main challenge in service projects for project team professionals remains to explain the technical expertise, knowledge and the value gained to the customer if the technicalities are not understood by the customer well. This is because the customer evaluates the value of the service project by what they can observe, such as interactions with the professionals if they do not have the technical knowledge themselves.

2.1.2 Marine Projects

Tvedten & Bauer (2022) explains that ships are one of the biggest machines on Earth and ships are delivering globally most of our traded goods. There are projections for expected increase in sea trade by 2050 range from 25% even up to 180%. To consider this, the transportation of traded goods by ships needs naturally a large amount of energy and today, with a few exceptional cases, ships rely mostly on fossil fuel and fossil fuel infrastructures.

Dixit & Tiwari (2021) explain that shipyards have a certain product range in regards of ship types, including several kinds of ships, like bulk carriers, barges, or tankers. Therefore, each newbuilding project is unique based on the vessel type and customer's specifications. Concurrently, the repetitiveness of same repetitive tasks and activities on similar vessel projects contributes to learning knowledge of how to do things better.

Lagemann, Lindstad, Fagerholt, Rialland & Erikstad (2022) explain that alternative fuels and fuel-flexibility of vessels are seen as promising solutions for achieving reductions of greenhouse gas (GHG) in the whole shipping industry. As IMO adopted strategic plan to align temperature goals of the Paris Agreement to reduce those GHG emissions in the international shipping industry. The first step to reduce the carbon density of ships can be achieved through further reductions of energy efficiency design index (EEDI) of ships. Secondly, carbon dioxide emissions by transport work should be reduced by 40% by 2030 and pursue 70% reduction by 2050 compared to 2008. Thirdly, the strive to peak international shipping industry's GHG emissions as soon as possible in order to reduce GHG emissions 50% by 2050 compared to 2008. Therefore, a newbuild ship designer has a large uncertainty associated with different factors in regards for aiming for a competitive ship.

Lagemann et al (2022) continues to explain that studies so far have mostly focused on technical details of emission reduction measure, alternative fuels, timing of air emission option. Alternative fuels have been estimated as the highest abatement potential and are very interesting to ship designers. Industry examples have already been shown as retrofits of The Spirit of British Columbia from very low sulphur fuel oil (VLSFO) to liquefied natural gas (LNG), and the conversion of Stena Germanica from heavy fuel oil (HFO) to Methanol. Methanol and Ammonia have shown significant interest for alternative fuels. Ship owners are therefore today in a conflict to find a cost-effective and long-term solution for their ships to stay competitive and within regulations.

Retrofit costs $C^{K}_{s'st}$ in [mUSD]. Green: low cost, red: high cost.

from/to	VLSFO ship	LNG ship	LPG ship	Methanol ship	Ammonia ship	LH2 ship
VLSFO ship	0.0	9.3	9.3	6.0	15.0	24.6
LNG ship	3.5	0.0	2.5	6.7	11.3	22.6
LPG ship	4.0	7.8	0.0	6.7	8.2	22.6
Methanol ship	1.0	9.3	9.3	0.0	15.0	24.6
Ammonia ship					0.0	22.6
LH2 ship					6.7	0.0

Picture 1 Estimated retrofit costs of conversions (Lagemann, Lindstad, Fagerholt, Rialland & Erikstad (2022)).

Lagemann et all (2022) do summarise that, for ship owner to decide to which kind of ship to put their money in, is very much affected by the desired emission level. Instead of building a new liquid hydrogen (LH2) or ammonia powered ship, which cannot yet be operated, the ship owners should think of service project retrofitting the vessel from example to LNG to ammonia when the development has become attractive as experience with ammonia and hydrogen as marine fuels are still to be built up. Retrofitting an existing vessel into different power-solution would require also changes to storage systems and their costs.

2.2 Planning of a Project

As explained by Dvir, Tzvi, & Shenrar (2003), there is a heavy desire and need to invest in project management team’s procedures and processes in order to support the planning of a project. Planning naturally reduces the uncertainty and increases therefore the chances of successful project delivery. However, a certain level of planning is done in all projects, as there has been advancement in computerized planning tools and a boom in training of project management.

Chang, Wiewiora & Liu (2021) explains that projects usually operate under major pressure to delivery products or services, and this leaves often very few resources to reflect lessons learned after a project completion, that could promote innovation or contribute long-term organizational success. Significant positive relationship can be found between the time and effort invested in planning of the project goals, functional requirements, product technical specifications and project success. However, these project learning activities are crucial to project teams and overall organizational success when they are effectively given to other projects or across organization as they help to avoid repeating same mistakes.

As explained by Dvir, Tzvi & Shenrar (2003), typically project manager’s goal is described to bring project to a completion in a certain time, within a certain cost plan budget, and to finally meet the planned performance and end-product goals. Therefore, it is usually seen view that the project manager’s task is always clear and well defined in advance to meet the performance or end-product goals.

According to Picciotto (2020) the linear concepts of “traditional” project management are being challenged now as it has been recognized that traditional project management practices are not really adapting to operational environments shaken by volumes of volatility and uncertainty. In these changing situations, a precise planning is often ineffective due to unrelatability in unstable systems and environment. Deviations from the initial project plans set and re-thinking project boundaries out of the box are frequently needed in order to handle changes and overcome unexpected challenges in time, cost-effectively. Therefore, project management theories are being “rethought” to broaden scope of project management.

According to Dvir et al (2003), at project initial stage it's often very difficult or even not possible to know beforehand every activity that should be completed out in order to successfully deliver the project, and to know exactly what the specific costs and duration parameters are. Even more difficult is to forecast activity which relies on outcome of earlier activity, and for that reason planning might not even be helpful at all.

Svejvig & Andersen (2015) explains that "traditional" project management is more like a mindset of viewing project management as a task-oriented execution of a list given to the project manager. The "re-thinking" of project management reflects this with a broader view of the whole concept involving a holistic perspective. The difference between the traditional view of project management and current trends is that projects are considered as more like temporary organizations instead of task lists. The new trend of project management involves lessons learned during the project and mindset of adaptive project management. However, the relationship between "old" and "new" project management views should not be excluding each other but on the contrary combine old views and new insights.

Dvir et al (2003) conclude that the importance initial planning stage of the project should be emphasized to properly defined targets and requirements. In order to successfully deliver the project, the end-user or customer should be involved in the planning process as much is needed. Customer involvement is very important in the phase of the technical specification of the product is important until the final specifications are agreed and locked. The involvement of the customer should start already at the initial phase and continue until the successful ending of the project. Project manager should be the key responsible of communicating the plans and the face of the organization to the customer.

2.2.1 Work Breakdown Structure in Projects

According to Dori & Sharon (2012), a Work Breakdown Structure is an extensively used planning method in projects. A Work Breakdown Structure is basically a hierarchical structure decomposition of a project into smaller activities into detailed level of work levels. These working levels represent activities and milestones of the project. Therefore, Work Breakdown Structure is reflecting the total scope of work to be done in the project. In order to serve this purpose, the Work Breakdown Structure levels is the source of project cost estimations, milestone scheduling and risk mitigation. Work Breakdown Structure should be used as to represent the deliverables of the project and needs to be based on the planned outputs throughout the project. This kind of approach produces a project plan throughout the project.

According to Aramvareekul & Seider (2006) cost measurements can be followed up on the Work Breakdown Structure levels, which is considered the foundation of project plan and where all the work that needs to be done in the project is listed.

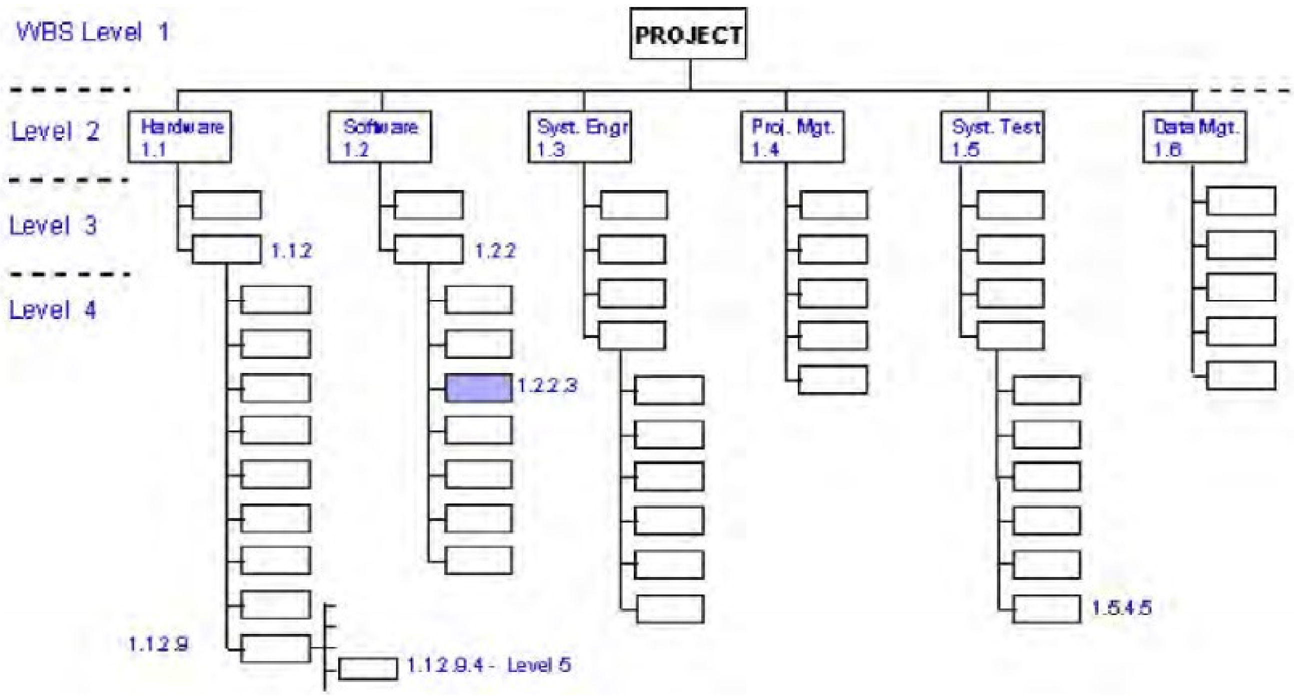


Figure 2 Work Breakdown Structure Hierarchy (Aramvareekul & Seider (2006)).

2.2.2 Cost Planning of a Project

Siami-Irdemoosa, Dindarloo & Sharifzadeh (2015) explain that cost planning & estimation of the project is done on the Work Breakdown Structure levels. Here in Work Breakdown Structure the costs of the project are allocated in the lowest-level components of the Work Breakdown Structures where those specific costs can be monitored and controlled separately.

According to Aramvareekul & Seider (2006), the main idea acknowledged of good project management is to make sure that projects will be delivered and completed on time, by the cost plan and while meeting customer's expectations. The management of these projects requires planning, scheduling, risk analysis, quality control and solution activities. Very often projects do involve people with different responsibilities and expectations from many different teams within and outside the organization as well. Even though every group has different goals, the main project goal for all is to deliver the project successfully. As projects are often of unique nature, the specific planning of activities remains challenging.

Aramvareekul & Seider (2006) continues to explain that project managers are able follow up the costs on the project with measurements such as planned cost of total work scheduled to be performed (BCWS), planned cost to complete the work (BCWP), and actual costs of work performed (ACWP). Based on these measurements, the project's ability to stay on the cost plan can be measured.

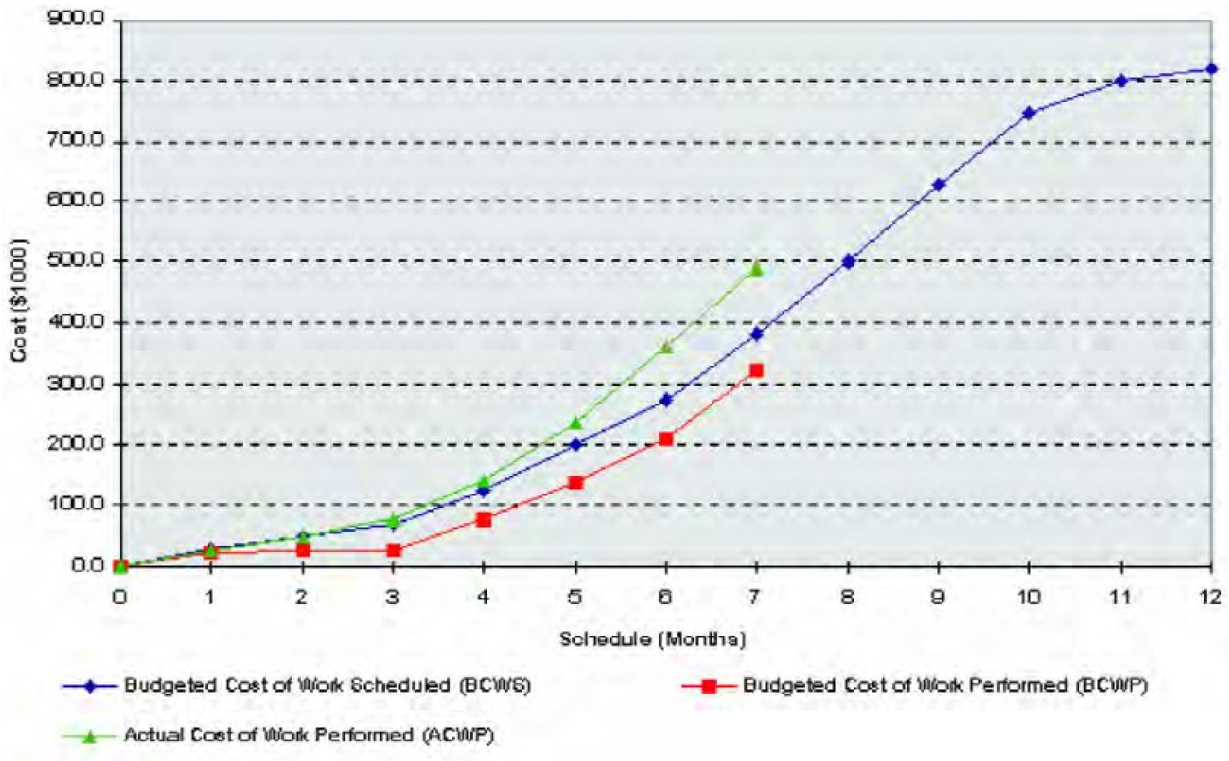


Figure 3 Cost Plan Followup (Aramvareekul & Seider (2006)).

2.2.3 Risks in Projects

According to Aramvareekul & Seider (2006) risk evaluation is an important step in project planning and project controlling in order to improve overall project decision-making and to reduce uncertainty. This is very important as in case risk is not managed, projects might be in serious problems with unplanned tasks which could result in negative cost overruns, delays in the schedule, issues with quality or in worst case scenario a total failure to meet project's objectives. These project risks are often grouped to three different categories: external risks, internal risks, and project specific risks. Project specific risks are characteristics which occur only for that specific type of project.

Kwon & Kang (2018) explains that risk is a crucial factor for a project in estimating project costs. The project managers still face the challenges of delivering projects within given cost plans but without the exceptional tools to deal with these unidentified risks. The cost is the most crucial parameter within the standard success criteria of cost, schedule, and performance targets when it comes down to management of projects.

Kwon & Kang (2018) continues to explain that while project managers have been managing projects to meet these very important targets, researchers have been trying to identify and find those root causes of negative cost overruns and develop accurate cost plan estimation methods to solve these problems. Previous research has found out that risks are one of the biggest reasons for negative cost overruns, therefore various cost plan estimation methods have been developed to estimate project reserves against risks. However, these methods are presented to only identified risks only and therefore projects still may suffer from negative cost overruns. Therefore, more accurate cost plan estimations would reduce surprising risks and thus negative cost overruns.

As explained by Kwon & Kang (2018), there is two categories of reserves in the cost plan, reserve for contingency for the identified risks and a reserve for management for unidentified risks. These two cost reserves in the cost plan are used for responsive actions. The contingency reserve is reserved against known risks and management reserve is to cover unknown risks which may include residual and secondary risks beyond the identified risks.

2.3 Estimation of Project Costs

As explained by Kwon & Kang (2018), the project cost plan should be estimated approximately to the actual costs in order to minimize cost differences. One of the successful ways in performing to improve accuracy of a project cost plan is to move uncertain scopes of work to certain scopes by analysing quantifying the uncertainty. This could be reserves in the cost plan against uncertainty. The reserve can be divided in the cost plan into adaptive action and preventive action by analysing the risks. Adaptive action means what is uncertain scope and preventive action becomes the certain scope. Therefore, the preventive action against risk should be then transferred to the Work Breakdown Structure.

Kwon & Kang (2018) continue, that the actual costs of a cost plan should be recorded for each project separately, and by analysing and evaluating the variance between initial cost plans and actual costs of each project are essential to improving project cost management. Analysing and evaluating variances between cost plans becomes very difficult if there are no cost plan estimations done for projects.

2.3.1 Forecasting of Project Costs

According to Iranmanesh & Zarezadeh (2008), one specific important problem for project manager’s team is accurate estimation of time and cost of work in completion for a project. Earned Value Management System (EVMS) is a known method with wide application for project cost forecasting. It is a systematic approach for measurement of project progress and calculate earned value & earned actual value and is a valuable management tool for project managers when estimating costs. The problem with estimating its completed cost is however crucial factor.

Iranmanesh & Zarezadeh (2008) explains that project performance is measured in EVMS by determining the cost planed cost of the work performed and comparing it with the actual cost of the work performed, that is used to estimate actual cost using Artificial Neural Network (ANN) method. By taking historical data of such projects with two cost plan values, it is possible to calculate the monthly or averaged index data, regression (non-linear & linear regression analysis) and forecast.

As Iranmanesh & Zarezadeh (2008) continues, forecasting with a neural network method involves two different steps, training and learning. When a training dataset is available to the developer, given with historical data containing both input and the output of the desired object. These are presented to the network. The input selections for neural network training is heavily influencing the success rate of training. Therefore, in the learning process of the neural network constructs a mapping of input-outputs, which is adjusting the different weights and biases of the iteration based on the minimizations of error measures between output produces and the desired outputs. Learning is therefore entailing an optimization process itself.

As explained by Iranmanesh & Zarezadeh (2008), comparison between actual and forecasted cost data show performance with Mean Absolute Percentage Error (MAPE). This can be used to help evaluating the accuracy forecasted of cost plan.

2.4 Business Intelligence

According to Ballard, C., Farrell, D., Gupta, A., Mazuela, C. & Vohnik, S. (2006), the survival of a company depends in today's very competitive world on how quickly they are able to recognize the changes in business dynamics and challenges, and how rapidly they are able to respond correctly. The key to succeed in any business is information and how they use that information. Business intelligence is therefore all about information and the home for storing this information is the enterprise data warehouse.

Ballard et al (2006) continue to explain, that companies collect significant amount of data and therefore need the ability to transform the raw data into actionable information. Business intelligence helps the company to gather and utilize this information as a competitive advantage. The competitive advantage is created by analysing historical business data, which is used to understand market situation, analyse competitors, business trends, strengths, weaknesses.

2.4.1 Data Modelling

Fisher, Watson, Escrig, Witt, Porcu, Bacon, Rigley and Gomes (2020) explain that models translate theories into mathematics. Once a model is built, the model can be used to help with making of decisions, developing scientific understanding and even to communicate knowledge and perform forecasts. Empirical data models are basically mathematical equations derived from data analysis and rely on the sufficiency of data quality and quantity. These capabilities of empirical data models have expanded recently to fields in machine learning and computational intelligence because of data-driven modelling. Machine learning can be used in development of algorithms which can access to read and learn from data. Data-driven models can be used to train algorithms such as linear regression & ANN on data.

Fisher et al (2020) continues to explain that data-driven algorithms have existed before, but their use has been limited due to constraints in the quality of data. The data previously available has been of poor quality or not stored in useable format for data-driven algorithms.

Ballard et al (2006) explains that dimensional data modelling focuses on the business, and it gives an improved possibility and capability to visualize abstract remarks business analysts may have. In a dimensional model, the data structure can be easily understood and navigated in order to utilize the data.

Ballard et al (2006) continue that model is basically an abstraction and reflection of a real world. Modelling helps us to visualize things we cannot see, and the same applies for data modelling. Therefore, without a data model, for a developer it would be very difficult to organize the structure of data and data contents in the data warehouse.

2.4.2 Dimensional Modelling

According to Ballard et al (2006), dimensional models are used to overcome performance issues of large queries in data warehouse. Dimensional model is often more commonly named a *star schema*. Star schema is well-known in data warehousing as it gives better query performance, and it is easy to understand. The heart of a star schema is often a *fact table*, which is a large data table of facts. The numerous surrounding tables around the star contain descriptive data are called *dimensions*. When the data model is created, it resembles a star.

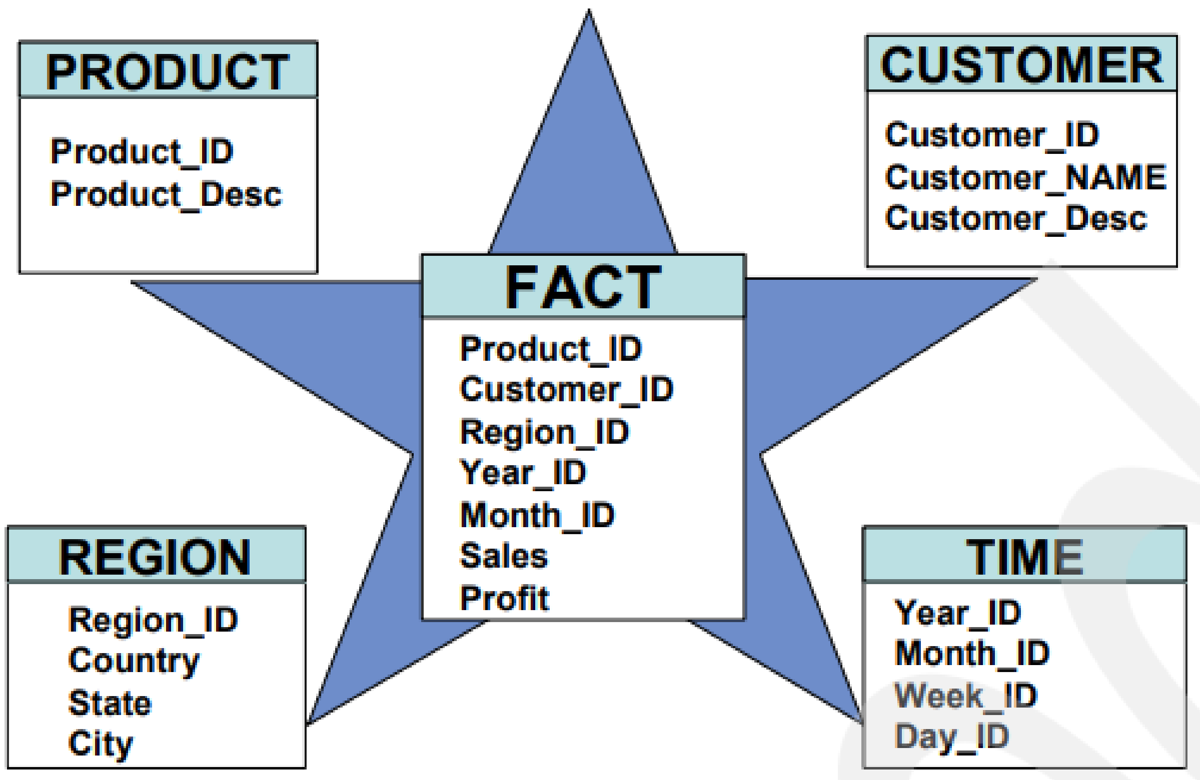


Figure 4 A star schema or a dimensional model (Ballard, C., Farrell, D., Gupta, A., Mazuela, C. & Vohnik, S. (2006)).

As Ballard et al (2006) explains, the fact table contains mainly keys which are associated with dimensional tables. Compared to the dimensional tables, fact tables consist of only few columns but a large number of rows. Dimensional tables therefore contain information about the facts. These are descriptive details about the numerical values in fact tables such as country name, product description or customer name.

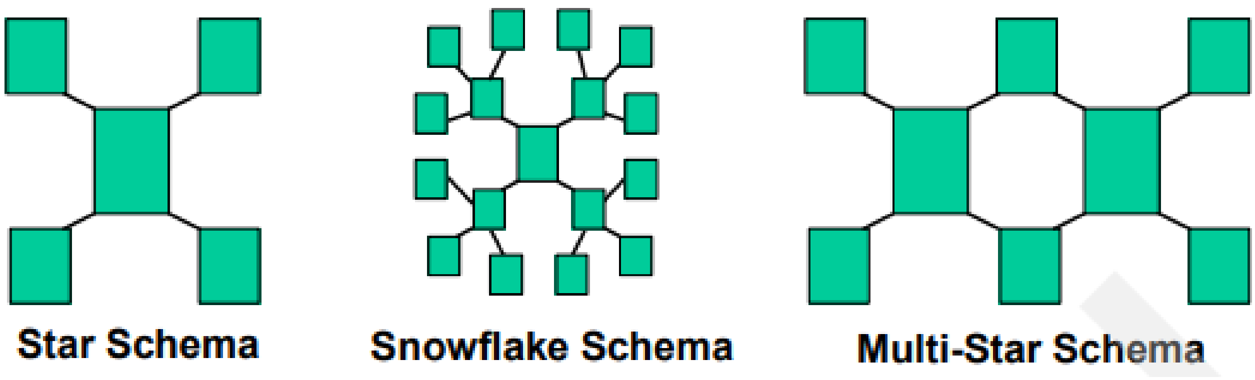


Figure 5 Types of dimensional models (Ballard, C., Farrell, D., Gupta, A., Mazuela, C. & Vohnik, S. (2006)).

As Ballard et al (2006) explains, as the data model keeps growing, the star schema may grow as well into snowflake schema or multi-star schema. These are basically expansions of the original star schema. Multi-star model is also a dimensional model, but it contains many different fact tables linked and joined together through dimension tables.

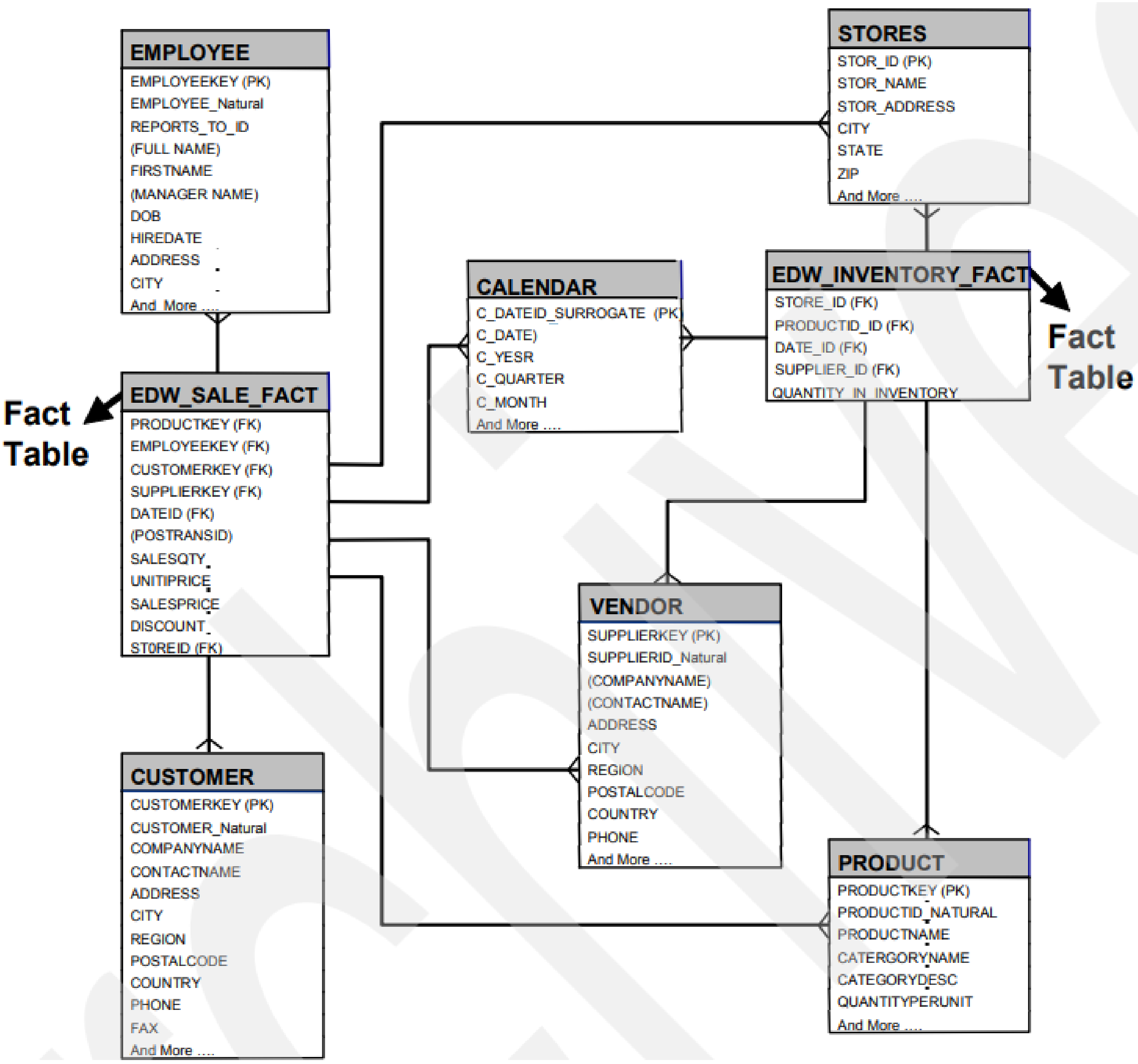


Figure 6 A multi-star model consists of multiple fact tables (Ballard, C., Farrell, D., Gupta, A., Mazuela, C. & Vohnik, S. (2006)).

2.4.2.1 Defining the Data-driven Model

As Fisher et al (2020) explains, there are considerations which should be made before creating a data-driven model and before data is collected or model built. The developer should consider defining the data model's goal, understand what is exactly required from the data model and what are the considerations of data. Developer should consider what data has been already collected, which data still needs to be collected, how much overall data is needed and how and when the data should be collected.

According to Fisher et al (2020) development of a data-driven model is composed in total to six steps: These steps are defined as understanding of business, understanding of data, preparation of data, modelling, evaluation & deployment. Business & data understanding steps are focusing on defining data-driven model understanding, business objectives and how to collect and explore these from the data. Data preparation step is important milestone as it contains the selection, cleaning, and data transformation in order to make formats to the data correctly for next phase which is modelling. Modelling step consists of various modelling techniques. Evaluation step is to evaluate how to use results gained from obtained models. Deployment step focuses on utilizing these results to benefit the business needs. Overall, the development process of a data-driven model is an iterative process as developer gains knowledge continuously with the process, which may lead then to redefining objectives and approaches.

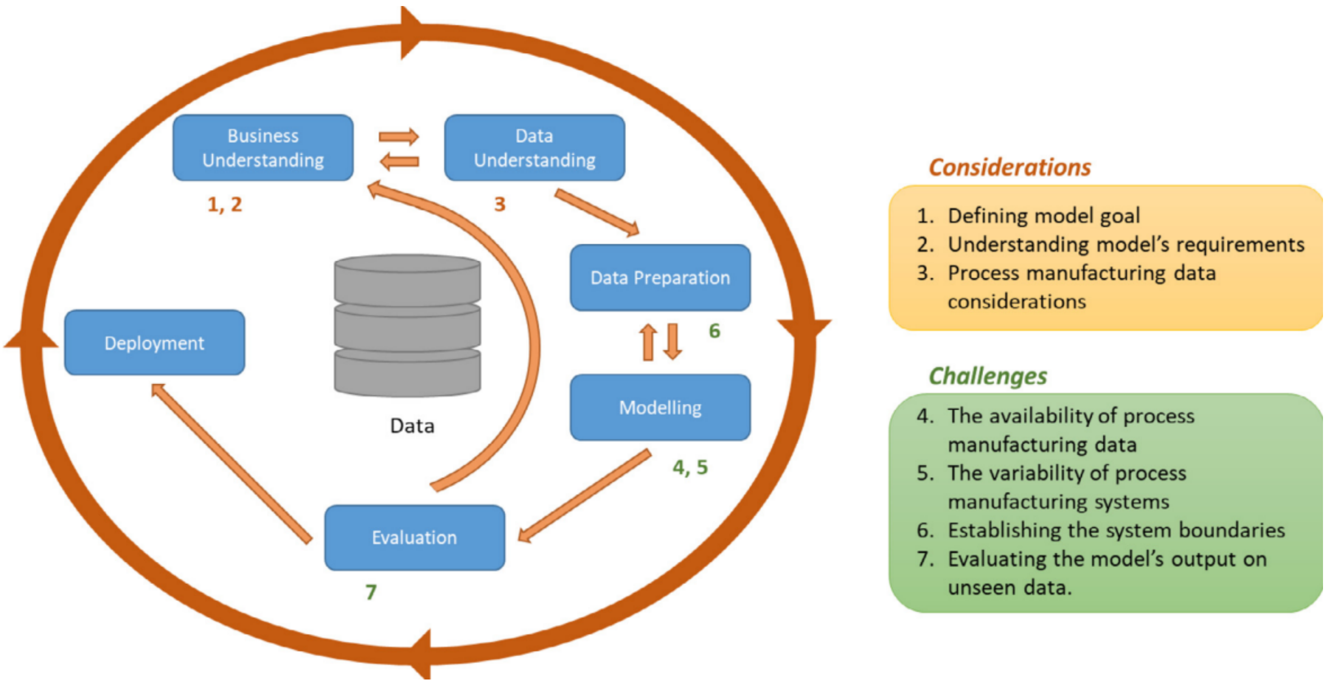


Figure 7 Considerations for development of data-driven model (Fisher, Watson, Escrig, Witt, Porcu, Bacon, Rigley and Gomes (2020)).

2.4.3 Data Warehouse

As Ballard et al (2006) explains data modelling is important, as it specifies the data structure which may impact all aspects of usage of data. The data modelling therefore may have a major impact on performance, especially with warehousing of data. The data warehouse is the primary structural element in business intelligence.

Soler, Trujillo, Fernández-Medina & Piattini (2008) also explain the importance of data warehouses. Data warehouses are the core of decision support systems as they provide the critical business information and therefore improve the process of making strategic decisions. The data warehouses are based on multi-dimensional modelling where information is structured into facts and dimensions.

Ballard et al (2006) argue that *enterprise data warehouse* is a data warehouse which is constructed and designed based on the business needs, to support all business requirements in a fully integrated data warehousing environment which has a high access to data and usage across the organization. The enterprise data warehouse stores data for decision-support across the organization.

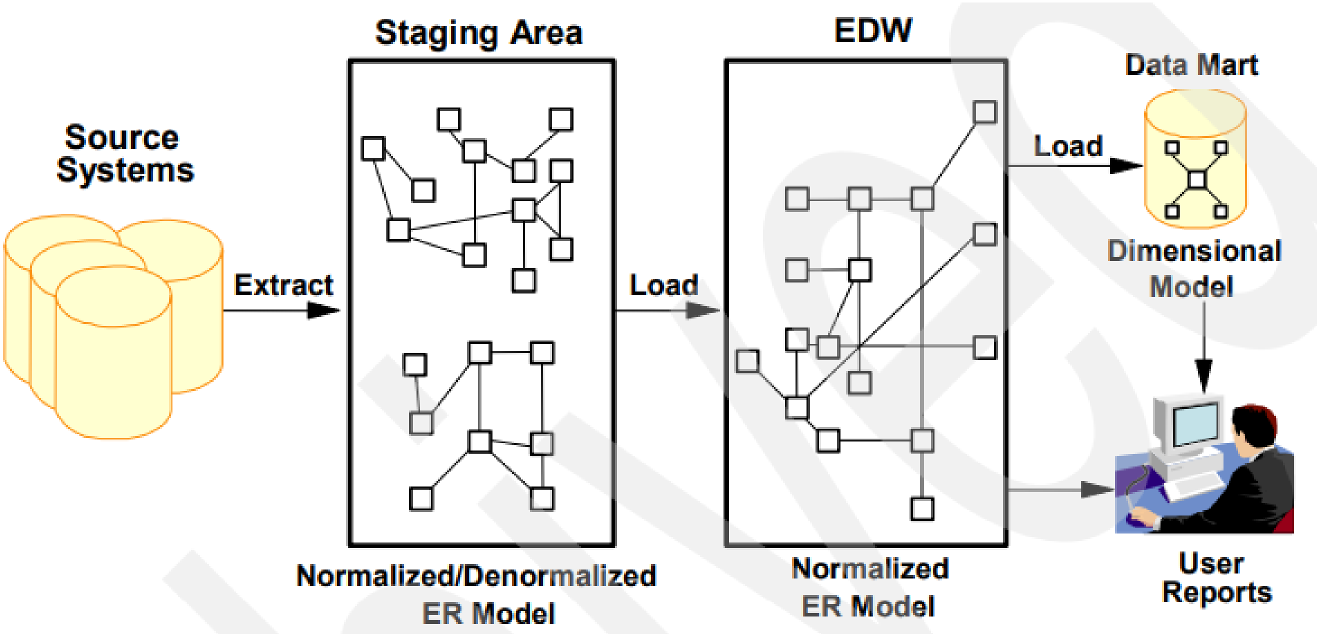


Figure 8 Enterprise data warehouse environment (Ballard, C., Farrell, D., Gupta, A., Mazuela, C. & Vohnik, S. (2006)).

As explained by Ballard et al (2006) in an enterprise data warehouse environment, the data is extracted from the source systems, which are the operational databases of the organization, into the data staging area, where the data is prepared by extracting and transforming and loaded (ETL) to the enterprise data warehouse. The report developer can then connect and utilize the data in the enterprise data warehouse for the business needs in reporting.

2.5 Data Analysis

According to Ballard et al (2006), to answer multidimensional questions where multiple dimension tables are needed, multidimensional analysis methods are used. As the data is categorized by different dimensions, the business-orientated users understand the answers easier. Dimensions can be hierarchies such as regions, stores, or departments. Multidimensional analysis will enable users to analyse large number of independent factors involved in the business case and visualize data in complex relationships. The data can be shown in lower levels of details or shown in higher hierarchy levels of summarization as per business needs to get insights.

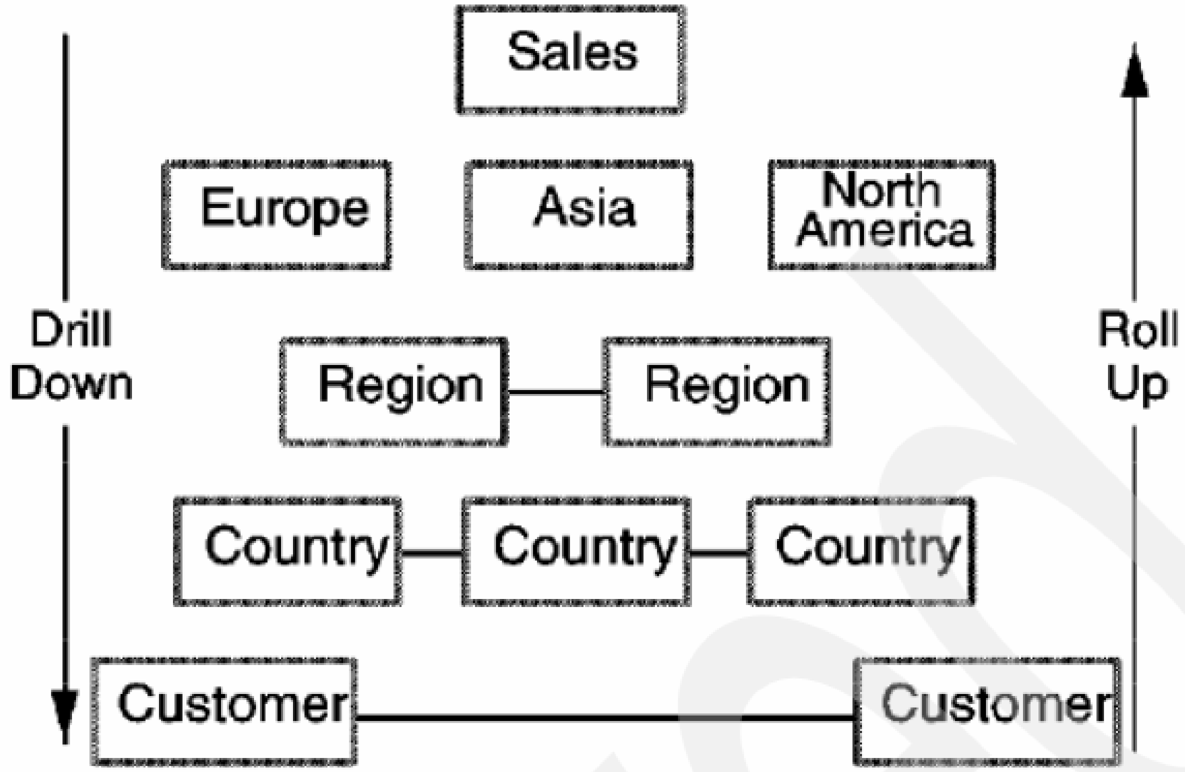


Figure 9 Drill down analysis (Ballard, C., Farrell, D., Gupta, A., Mazuela, C. & Vohnik, S. (2006)).

As Ballard et al (2006), explains statistical data analysis is used to detect unusual data patterns and to apply mathematical and statistical modelling techniques to try and explain those patterns. These models are usable in order to forecast and forecast results, and techniques to use in statistical data analysis can be such as nonlinear and linear analysis and regression analysis. Utilizing these techniques to right dataset may lead to discovering previously not yet known information from the data and uncovering unknown business facts.

3 Method

This chapter explains the research methods used in conducting this research. The chapter aims to give insight to the reader and make the replication of the research possible in a similar case study. The prerequisites of the data gathering process is explained more in depth, and the reasoning for choosing different angles is opened.

3.1 Research Methods

This research is a qualitative and quantitative research which is examining data gathered from dataset and earlier literature on the related topics. The purpose of the quantitative research on the thesis is to find out common characteristics to describe causes to phenomenon and relationships. The qualitative research of the thesis strives to find statistically significant results to describe the phenomenon and relationships.

3.2 Research Process

The data for almost 2500 marine service projects were gathered from Enterprise Data Warehouse. Number of around 2500 projects were chosen in order to produce enough diverse set of data to the research. The dataset gathered consisted of main project data, financial data, and descriptive project data.

The main principle in gathering of the desired dataset was to restrict the search results based on the metadata of each project to limit the results to the desired projects only. This filtering of dataset was done by Structured Query Language (SQL).

After gathering the desired dataset, the data was analysed to find out insights and relationships about the projects. The statistical analysis in the research was conducted based on the gathered dataset and a regressor model was created to forecast new values. These steps are explained in the next chapters more closely.

3.2.1 Research Data

The organization's Enterprise Data Warehouse contains data from Enterprise Resource Planning system and other sources but for this research, only data from the Enterprise Resource Planning system was used. The tables need for this research consisted of fact and dimensional tables. The most important tables included project data, financial data, and different dimensional data such as customer, country, installation segment, installation type, products involved in the project and nominated project manager.

The data tables were queried from the Enterprise Data Warehouse in Microsoft PowerBI cloud workspace using dataflow. SQL was used to filter on the desired results and in some queries, tables were combined with SQL. The visualization of the data was also created with Microsoft PowerBI desktop.

Picture 2 SQL queries in Microsoft PowerBI Dataflow.

The queries downloaded to the Microsoft PowerBI cloud workspace. Here formatting, merging and calculations were made to the data tables. To keep the data up to date, the queries were split based the need for refresh frequency to slow & fast refresh rate of the dataflow as dataflows can be set to refresh on different schedules.

The data collected in the dataflows were linked to the PowerBI desktop -file where the data-driven multidimensional model and visualizations were created.

3.2.2 Data-driven Multidimensional Model

A multidimensional model was created in Microsoft PowerBI desktop file based on the tables queried with SQL from the Enterprise Data Warehouse in Microsoft PowerBI workspace's Dataflow. The queries were connected to the desktop -file from the PowerBI workspace. The multidimensional model created in the desktop is the heart of the report generation and in the centre of this model, is a main fact table, surrounded by dimensional tables and assisting fact tables.

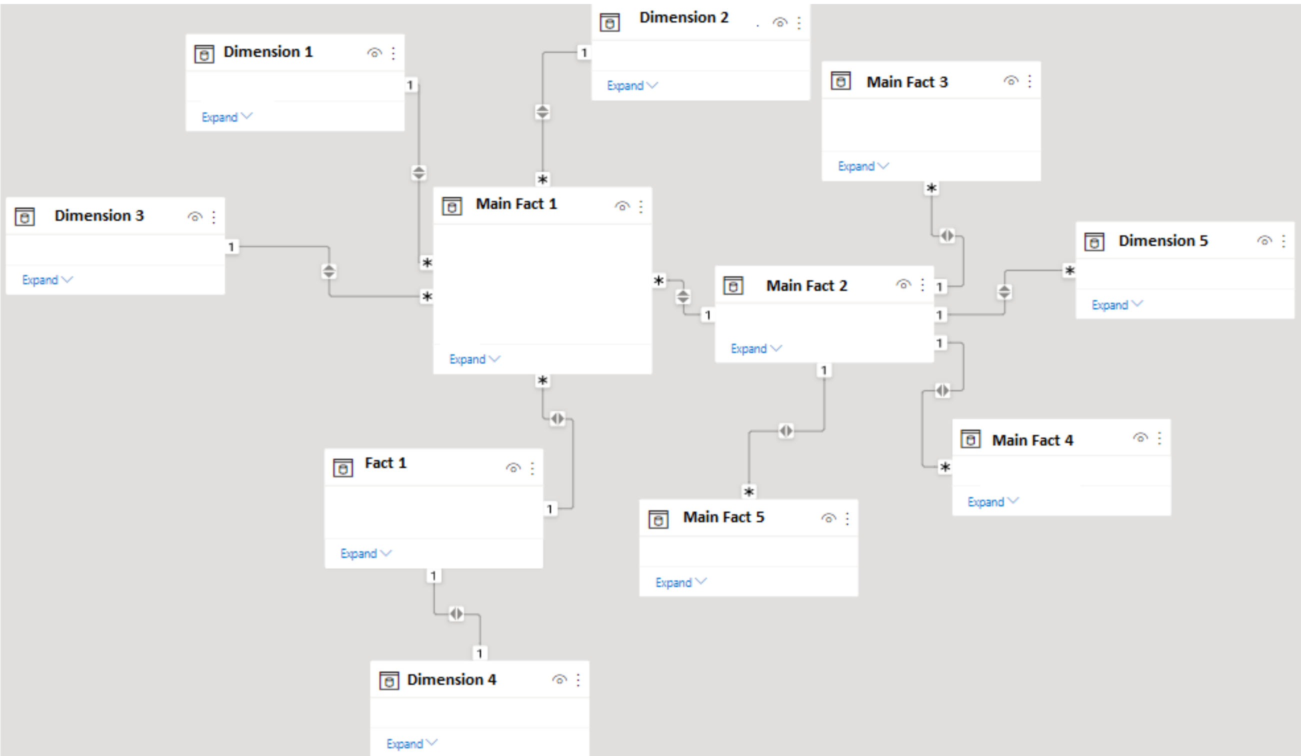


Figure 10 The multidimensional model including relationships between fact and dimensional data tables.

In the model the Main Fact 1 is a fact table and the main source tables of this research, consisting of project cost plans. The data table is combined with SQL to another fact table containing key project data and to five other dimensional tables where the key information is stored. This was needed in order to restrict values into one table for the regression model. The SQL query is then filtered to restrict the query to the desired scope.

There was a need to generate more data in the model, therefore the model chosen was multidimensional model consisting of multiple fact and dimensional tables. These data tables relate to each other with key values together using many-to-one, one-to-many and one-to-one relationships.

The other tables in the multidimensional model's schema are used to do comparison between project plan costs. Dimensional tables in the schema are translating the key values from the Main Fact tables into descriptive values.

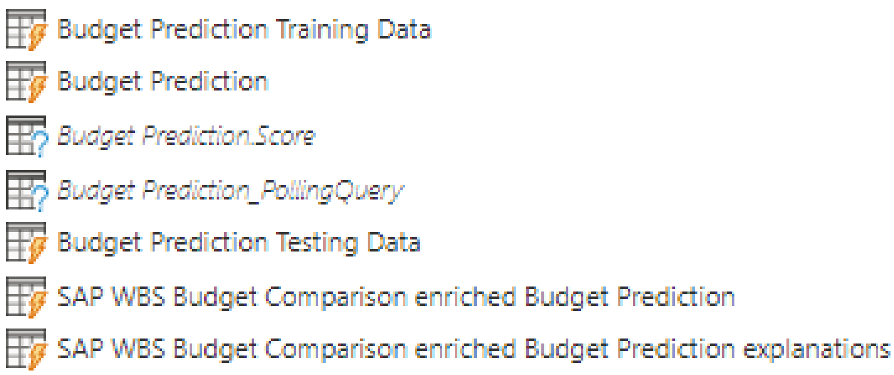
3.2.3 Regressor Model

The regressor model was created in the Microsoft PowerBi cloud software in the organizations workspace in order to analyse the dataset. The dataset was collected from the enterprise data warehouse with Microsoft PowerBI dataflows and the forecast was created there based on the gathered dataset.

Microsoft PowerBI dataflow utilizes automated Machine Learning capability in Azure Machine Learning to train the model. In order to create the regressor model, the dataset was combined into one query table consisting of all needed attributes of the project for analysis.

The key columns inside the query, *"main query",* where all queries were joined, 7 in total, were chosen to feed the machine learning model for training. These key columns consisted of descriptive project data characteristics and financial data.

The actual cost plan column consisting of values in Euro was set as target for forecast and key columns consisting of such as customer, customer country, installation segment, installation type, product type was set to contribute to the regression result.



Picture 3 Queries created by the Regressor Model in Microsoft PowerBI Dataflow.

The sample dataset consisted of total 2473 rows and training rows for the regressor analysis forecast were 1979. These training rows were used to train the regressor model in order to produce results for the live regressor model. These training rows were iterated for 25 times to find out suitable regressor estimator model.

The Pre-fitted Soft Voting Regressor estimator model was chosen by the machine learning model after the model training. This model was chosen as the model scored highest out of the 25 possible models, 0.92 quality points out of 1.00 in the training dataset. The performance of Pre-fitted Soft Voting Regressor model was measured by comparing the column containing the original cost plan value to the forecasted cost plan value by the Pre-fitted Soft Voting Regressor model. Pre-fitted Soft Voting Regressor is a machine learning regressor model which is predicting the class value with the highest probability, averaged over all the individual variables inserted.

Index	Score	Estimator Name	Estimation
1	0.76	Light GBM Regressor	LightGBMRegressor(min_data_in_leaf=20)
2	0.90	ElasticNet	ElasticNet(alpha=0.001, l1_ratio=1, normal-ize=False)
3	0.74	Light GBM Regressor	LightGBMRegressor(boosting_type=gbdt, colsample_bytree=0.9, learn-ing_rate=0.1579, max_bin=255, max_depth=9, min_data_in_leaf=0.0123, min_split_gain=0.10526, n_estimators=25, num_leaves=63, reg_alpha=1.05, reg_lambda=1.125, subsample=0.8, sub-sample_freq=4)
4	0.49	Random Forest Regres-sor	RandomForestRegressor(bootstrap=False, criterion=mse, max_features=sqrt, min_samples_leaf=0.0042, min_sam-ples_split=0.01073, n_estimators=50)
5	0.55	Random Forest Regres-sor	RandomForestRegressor(bootstrap=False, criterion=mse, max_features=0.7, min_samples_leaf=0.06102, min_sam-ples_split=0.01281, n_estimators=25)
6	0.47	Extra Trees Regressor	ExtraTreesRegressor(bootstrap=True, crite-rion=mse, max_features=0.5, min_sam-ples_leaf=0.00615, min_sam-ples_split=0.10734, n_estimators=400)

7	0.70	Extra Trees Regressor	ExtraTreesRegressor(bootstrap=False, criterion=mse, max_features=0.7, min_samples_leaf=0.0042, min_samples_split=0.03709, n_estimators=25)
8	0.91	ElasticNet	ElasticNet(alpha=0.001, l1_ratio=0.94789, normalize=False)
9	0.56	Random Forest Regressor	RandomForestRegressor(bootstrap=True, criterion=mse, max_features=0.8, min_samples_leaf=0.04163, min_samples_split=0.00183, n_estimators=400)
10	0.90	ElasticNet	ElasticNet(alpha=0.36905, l1_ratio=0.01, normalize=False)
11	0.37	Random Forest Regressor	RandomForestRegressor(bootstrap=True, criterion=mse, max_features=0.3, min_samples_leaf=0.06102, min_samples_split=0.00183, n_estimators=200)
12	0.91	ElasticNet	ElasticNet(alpha=0.05358, l1_ratio=0.68737, normalize=False)
13	0.60	Extra Trees Regressor	ExtraTreesRegressor(bootstrap=True, criterion=mse, max_features=0.8, min_samples_leaf=0.0042, min_samples_split=0.10734, n_estimators=10)
14	0.91	ElasticNet	ElasticNet(alpha=0.10616, l1_ratio=0.01, normalize=False)

15	0.56	DecisionTreeRegressor	DecisionTreeRegressor(criterion=mse, max_features=None, min_samples_leaf=0.03703, min_samples_split=0.0153, splitter=best)
16	0.47	Extra Trees Regressor	ExtraTreesRegressor(bootstrap=True, criterion=mse, max_features=0.8, min_samples_leaf=0.00615, min_samples_split=0.0153, n_estimators=25)
17	0.63	Light GBM Regressor	LightGBMRegressor(boosting_type=gbdt, colsample_bytree=0.5, learning_rate=0.14737, max_bin=63, max_depth=7, min_data_in_leaf=0.01719, min_split_gain=0.26316, n_estimators=50, num_leaves=3, reg_alpha=0.9, reg_lambda=0.075, subsample=0.6, subsample_freq=4)
18	0.73	Extra Trees Regressor	ExtraTreesRegressor(bootstrap=True, criterion=mse, max_features=None, min_samples_leaf=0.00347, min_samples_split=0.0009, n_estimators=100)
19	0.89	SGD Regressor	SGDRegressor(epsilon=0.09801, eta0=0.01, fit_intercept=True, l1_ratio=0.18367, learning_rate=constant, loss=squared_loss, n_iter=100, penalty=l2, power_t=0.2222, tol=1e-05)

20	-0.12	SGD Regressor	SGDRegressor(epsilon=0.0409, eta0=0.0001, fit_intercept=True, l1_ratio=0.55102, learning_rate=optimal, loss=huber, n_iter=1000, penalty=none, power_t=0.55556, tol=0.01)
21	0.91	ElasticNet	ElasticNet(alpha=0.42163, l1_ratio=0.06211, normalize=False)
22	0.89	ElasticNet	ElasticNet(alpha=0.21132, l1_ratio=0.32263, normalize=False)
23	0.91	ElasticNet	ElasticNet(alpha=0.26389, l1_ratio=0.01, normalize=False)
24	0.49	Random Forest Regressor	RandomForestRegressor(bootstrap=False, criterion=mae, max_features=None, min_samples_leaf=0.01938, min_samples_split=0.12814, n_estimators=200)
25	0.92	Pre-fitted Soft Voting Regressor	PreFittedSoftVotingRegressor(min_models=1, max_models=15, model_seed_threshold=0.05)

Table 1 Regressor Model qualities over iterations.

To get an indication on how the regressor model really performed, a Coefficient of Determination in comparison between predicted and actual values in column of original cost and forecasted cost value was measured, and the result was up to 94% in training dataset.

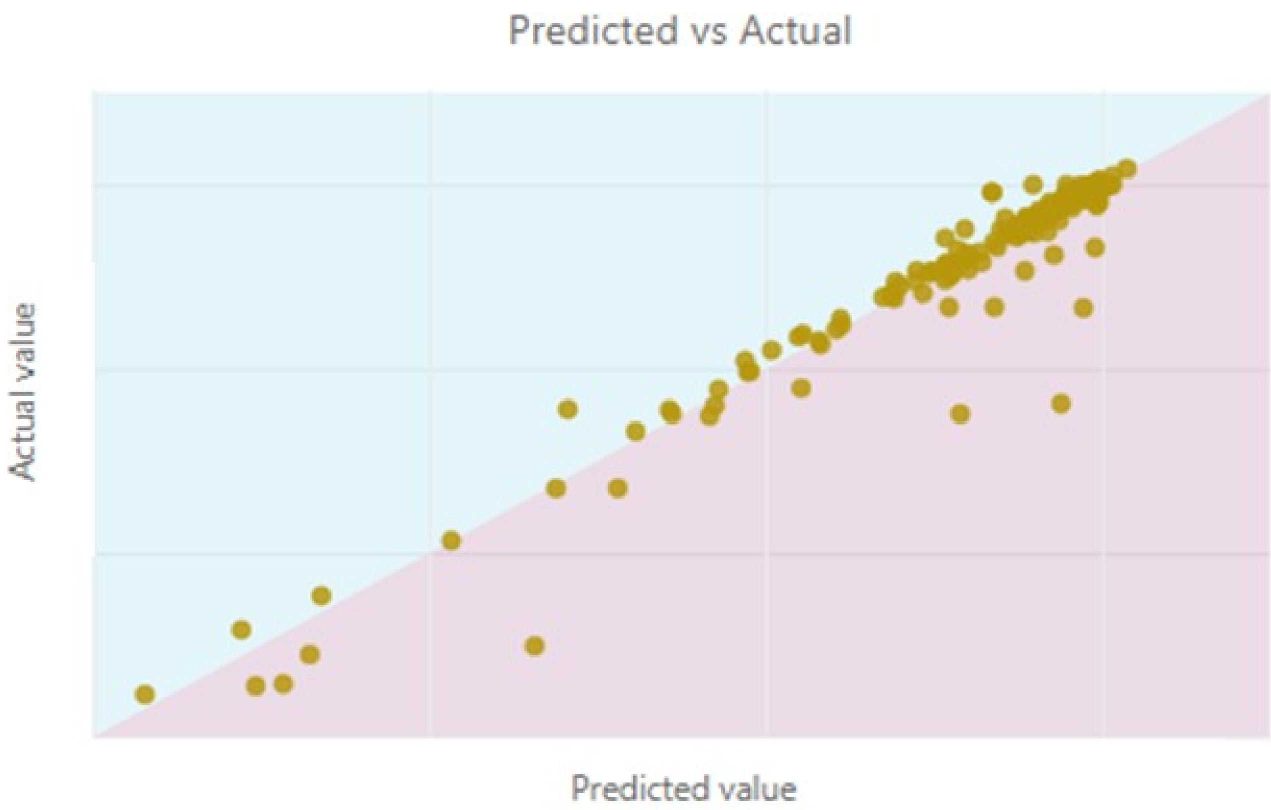


Figure 11 Model performance (Coefficient of Determination) of the Pre-fitted Soft Voting Regressor model.

However, in the Pre-fitted Soft Voting Regressor model's training dataset there were overestimates and underestimates. These overestimates and underastimates were evaluated on the residual percentage of error. By optimal situation these values should be close to 0%.

This result means there is a possibility that there will be projects with underestimates and overestimates of the cost plan in the actual dataset results, which should be taken into consideration. On average some of these overestimates of the cost plan could be up to +5% where some underestimates of the cost plan could be even down to -15%.

Figure 12 Residual error percentage of the Pre-fitted Soft Voting Regressor model.

3.3 Reliability of the Research

The good moral of scientific practice has been followed in this research. The steps and methods are listed where possible here for the reader to re-create the results similarly as the research. The main principles in this research should be able to be replicated if the similar data is available for the researcher.

The data selected for the study consisted of 2473 conventional marine service projects in the business sector and projects before and during the COVID-19 pandemic. Potentially, as in beginning of 2022, global price inflation of parts or services and war in Ukraine could heavily influence the results of this results or new research as all costs would rise greatly, which would make values greater. Therefore, this dataset gives the research results some additional divergence as in projects started in 2020 COVID-19 pandemic affected some personnel costs for example.

4 Results

This research focused and examined projects of a global organization performing customer delivery projects in marine services industry. These service projects are executed to provide services to variety of different products installed in different kind of installations. The vast majority of service projects (84,36%), project type 1, are upgrade work done to an existing product installed on the vessel.

Type	Projects
Project type 1	84.36%
Project type 2	7.26%
Project type 3	5.70%
Project type 4	2.68%
Total	**100.00%**

Figure 13 Most common service project types.

The research analysed the dataset consisting of main project data, dimensional data of the project and financial data. The dataset was analysed to identify characteristics of risky projects, and a Pre-fitted Soft Voting Regressor model was created to in order to produce forecast cost plan values and evaluate the accuracy of initial cost plans.

4.1 Analysis of Cost Overruns

The dataset contained clear cost deviations between initial project cost plan set by the sales team and actual cost plan by the project team. These deviations were either positive or negative cost overruns. In this research we mean with positive cost overrun a overrun which has less actual costs than initial costs. Negative cost overrun is therefore the oppositive with more actual costs than initial costs. There was no data of project where the initial project cost plan was exactly the same as actual cost plan.

The overall difference in the dataset when considering all characteristics, all segments between initial cost plan and actual cost plan were a negative -6,04%, resulting that in total, the projects initial costs were underestimated.

The projects in segment 9 consisted of the highest negative cost overruns with 1,38% and projects in segment 4 contained the highest positive cost overruns with 1,03%. There are the highest number of projects in Segment 9 (35,03% of all projects), while Segment 4 has the second lowest number of projects (2,07% of all projects).

Segment	%Initial	%Actual	Projects
Segment 1	0.14%	0.13%	0.93%
Segment 4	8.84%	7.81%	2.07%
Segment 5	25.37%	25.04%	14.30%
Segment 6	26.28%	25.69%	20.00%
Segment 7	12.09%	12.48%	16.99%
Segment 8	5.13%	5.33%	10.67%
Segment 9	22.14%	23.52%	35.03%
Total	**100.00%**	**100.00%**	**100.00%**

Figure 14 Installation Segments by percentage of projects.

Service projects conducted to vessel type 64 (9,41%), 1 (8,58%), 33 (8,17%) & 63 (6,83%) have clearly the highest amount of all projects. Vessel types 64, 33 & 6 are with positive cost overruns whereas vessel types 1 & 63 have negative cost overruns.

Installation Type	%Initial	%Actual	Projects
Vessel Type 64	2.78%	2.50%	9.41%
Vessel Type 1	5.43%	5.70%	8.58%
Vessel Type 33	14.99%	14.90%	8.17%
Vessel Type 63	4.88%	5.02%	6.83%
Vessel Type 6	3.25%	3.21%	5.38%

Figure 15 Vessel types by percentage of projects.

Service projects delivered to country name 11 (29,99%), 61 (7,45%), 1 (7,24%), 2 (3,93%), 99 (3,93%) have the highest number of all projects in the dataset. The Customer Country Name 11 is by far the most common customer country, but there are most negative cost overruns (-0,89%). Negative cost overruns are also in country name 1 (-0,45%) & 2 (-0,07%). Positive cost overruns are in country name 61 (+0,07%) and country name 99 (+0,06%).

Country	%Initial	%Actual	Projects
Country Name 11	28.21%	29.10%	29.99%
Country Name 61	3.66%	3.59%	7.45%
Country Name 1	10.20%	10.65%	7.24%
Country Name 2	3.94%	4.01%	3.93%
Country Name 99	3.44%	3.38%	3.93%

Figure 16 Customer countries by percentage of projects.

Product types 96 (12%), 12 (8,69%), 50 (6,72%), 48 (5,38%), 98 (3,31) in service projects are in most of all projects. Product type 96, 50, 48 & 98 have all positive cost overruns while product type 12 has negative cost overruns (-0,46%).

PRT	%Initial	%Actual	Projects
Product Type 96	4.41%	4.22%	12.00%
Product Type 12	22.96%	23.42%	8.69%
Product Type 50	1.45%	1.42%	6.72%
Product Type 48	0.38%	0.36%	5.38%
Product Type 98	6.59%	6.43%	3.31%

Figure 17 Product types by percentage of projects.

Top six project managers with the most projects also have differences in the initial and actual cost plans. PM 1 (10,13%) and PM 14 (7,86%) have clearly most of the projects, but PM 1 is clearly having negative cost overruns (-0,67%) more, and PM 14 has positive cost overruns (+0,13%).

Project Manager	%Initial	%Actual	Projects
PM 1	3.09%	3.76%	10.13%
PM 14	4.85%	4.72%	7.86%
PM 5	2.40%	2.66%	5.27%
PM 7	3.46%	3.58%	5.17%
PM 12	1.49%	1.59%	4.45%
PM 11	5.70%	5.64%	3.93%

Figure 18 Project Managers by percentage of projects.

All of these characteristics give some indication on which entities are related to the number of projects. By analysing these, we could identify that there are certainly some entities which tend to have negative cost overruns, which could mean that these are considered as risky characteristics of a service project delivery in the dataset.

4.2 Identification of Risky Project Characteristics

As the projects are executed to different kinds of installations, in variety of locations around the world and to different products, there is a great variety in project types as the environment is always different.

However, to pace it down, the projects are very similar to each other as in all of them something is manufactured and installed to the installation. As this is taken to consideration, the ability to identify a risky project relies on the expertise of the organization and different external factors.

By analysing the dataset, it was possible to identify segments where the most negative cost overruns between initial and actual cost plans. These segments were taken for further analysis. According to the analysis, the segments with most total costs are in segments 9, 5, 6 & 7.

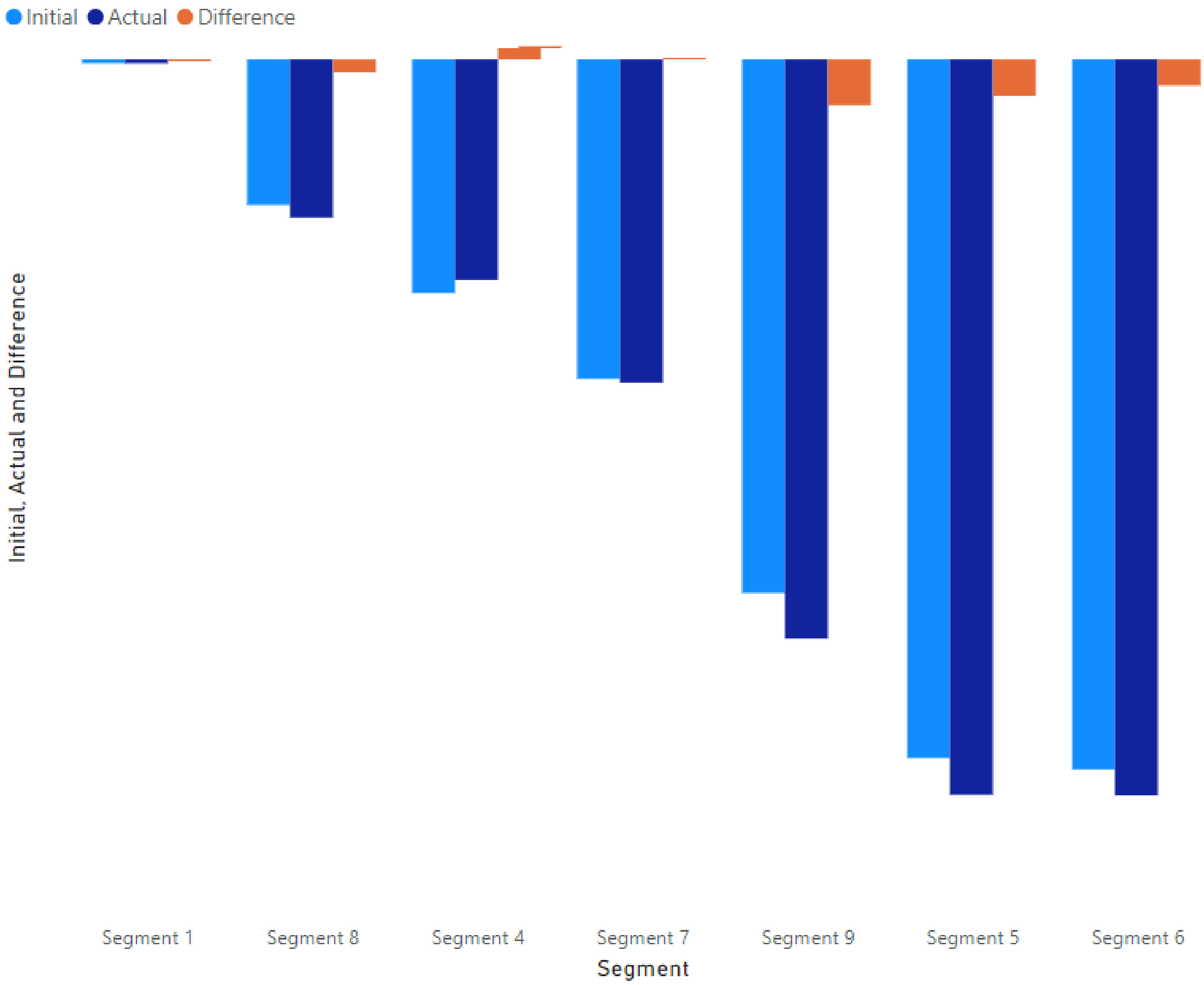


Figure 19 Cost plan differences per segment.

Taking a closer look at the segments and drilling through in the analysis line in Figure 20 by the deviations in cost plans, the vessel type 1 in segment 9 is collecting the highest negative cost deviations.

The characteristics in segment 9 does not clearly relate to segments 5, 6 or 7 attributes but the delivery of product type 12 is significantly present in segments 5 and 6 as well, there is also a clear match in segment 5 and segment 7 projects to country name 11 and delivering product type 12 to segments 5 and 6. These insights show that a service project to product type 12 could identified as a risky.

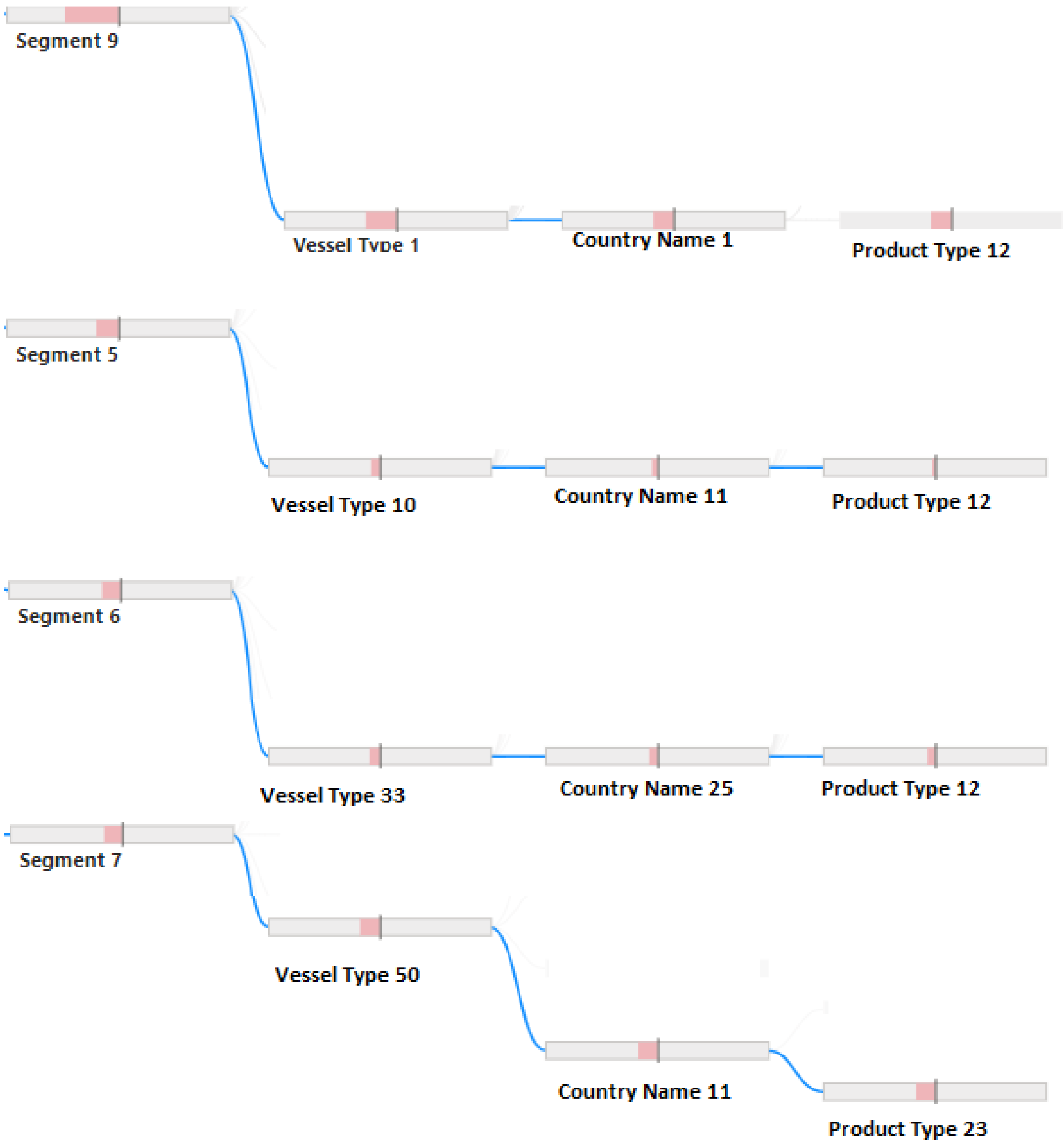


Figure 20 Drill-through of cost plan deviations in Segments 9, 5, 6 & 7.

Taking looking at negative cost differences in Figure 21, related to different segments, it can be noted that project delivered to country name 11 and country name 1 are on the top of the list. Country name 11 is by far collecting the most costs as country.

Drilling-down with the product type 12 in Figure 21, it can be clearly identified that it is linked to negative cost overruns. Segment 9, 5 & 6 can be noted down as top three with most costs, where segment 9 has the most costs once again.

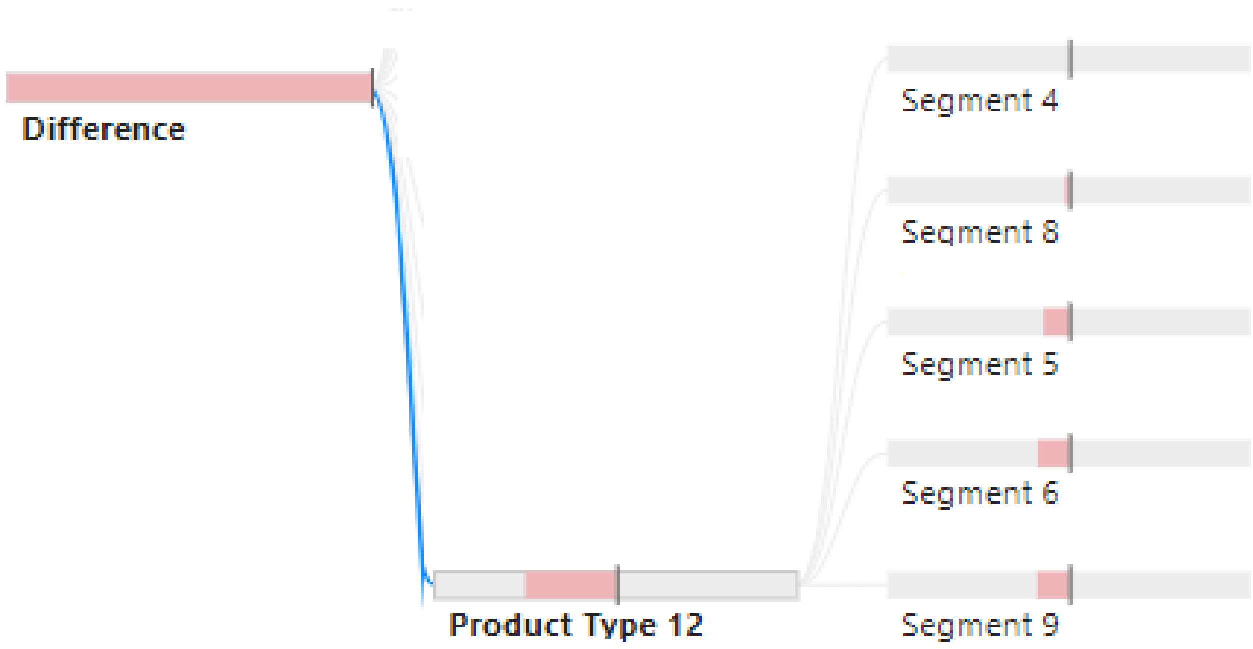


Figure 21 Product type 12 has highest cost differences in Segment 9.

By looking at highest cost differences by countries, it can be seen that the country name 11 has clearly the highest cost differences and therefore can be identified as highest risky characteristic in countries.

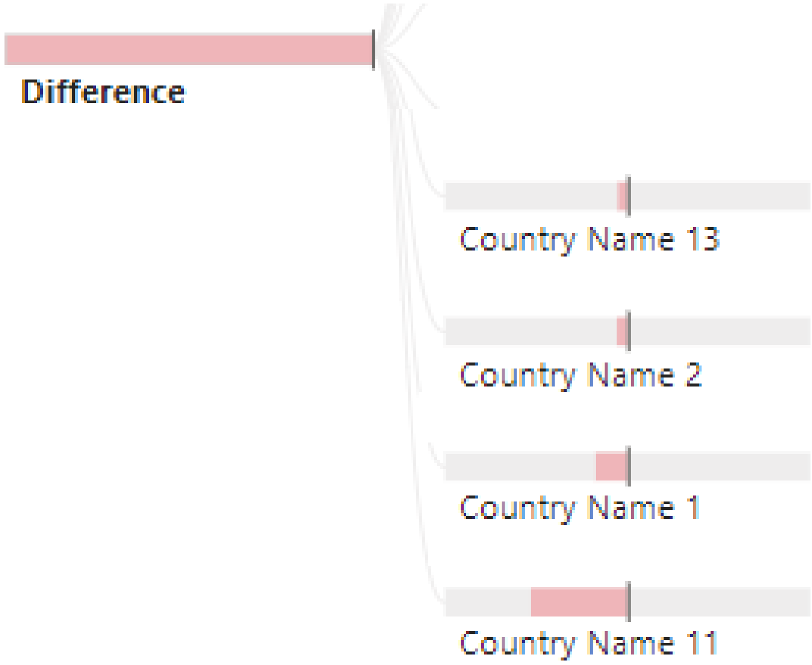


Figure 22 Highest cost differences by country.

The nominated project manager of the project can also have an effect of the project's risk level. By analysing the dataset, it can be noted down that project managers 1, 2, 3, 4 & 5 have significantly more costs in their project's actual cost plan compared to initial cost plan.

Project Manager	% Initial	%Actual	%Difference
PM 1	3.09%	3.76%	20.78%
PM 2	4.09%	4.44%	13.20%
PM 3	4.74%	5.00%	11.59%
PM 4	5.42%	5.63%	11.21%
PM 5	2.40%	2.66%	9.15%
PM 6	2.20%	2.39%	7.28%
PM 7	3.46%	3.58%	6.68%
PM 8	0.32%	0.50%	4.99%
PM 9	0.00%	0.19%	4.99%
PM 10	4.48%	4.49%	4.79%
PM 11	5.70%	5.64%	4.30%
PM 12	1.49%	1.59%	4.13%
PM 13	5.18%	5.11%	3.79%

Figure 23 Project Managers can influence project costs.

4.3 Actual Project Cost Forecasting

Forecasting actual project cost is a difficult task beforehand due to a lot of unknown elements in the project environment. Risky project charasteristics have been identified especialy in segments 9, 5, 6 and 7, whereas product type 12 seems to gather the most negative cost overruns, especially when delivering to country name 12. Risky project managers were identified as project managers 1, 2, 3, 4 and 5.

For this section a Pre-fitted Soft Voting Regressor model was created to produce and forecast cost plan values and to give insight if we are able to forecast the risky charasteristics beforehand.

In order to forecast unknown costs we need to take into consideration the dimensional data of the project, what kind of vessel is it in which vessel segment, where the project is executed, who is the owner of the vessel, who is in charge of the project execution and what is the product in the project but also how much costs were estimated to the project at initial stage before execution.

4.3.1 Forecasting with a Pre-fitted Soft Voting Regressor Model

Categorical data from the dataset was set as input to the Pre-fitted Soft Voting Regressor model to forecast the cost plans. Main input columns consisted of customer country, installation segment, installation type, product type related to the project, owner of the installation and project manager in charge of execution. Initial cost plan of the project was given as the first estimate.

As a first test of the Pre-fitted Soft Voting Regressor model, a scatter plot on overall level of the dataset was created to visualise any differences of the initial and forecasted values. This was done in order to see how the forecast is performing with the given dataset for the initial values which were known at the time of project creation.

The results in Figure 24 imply that the forecast of the Pre-fitted Soft Voting Regressor model is very much in linear line to the initial cost, which means that the Pre-fitted Soft Voting Regressor model seems to be linear regression.

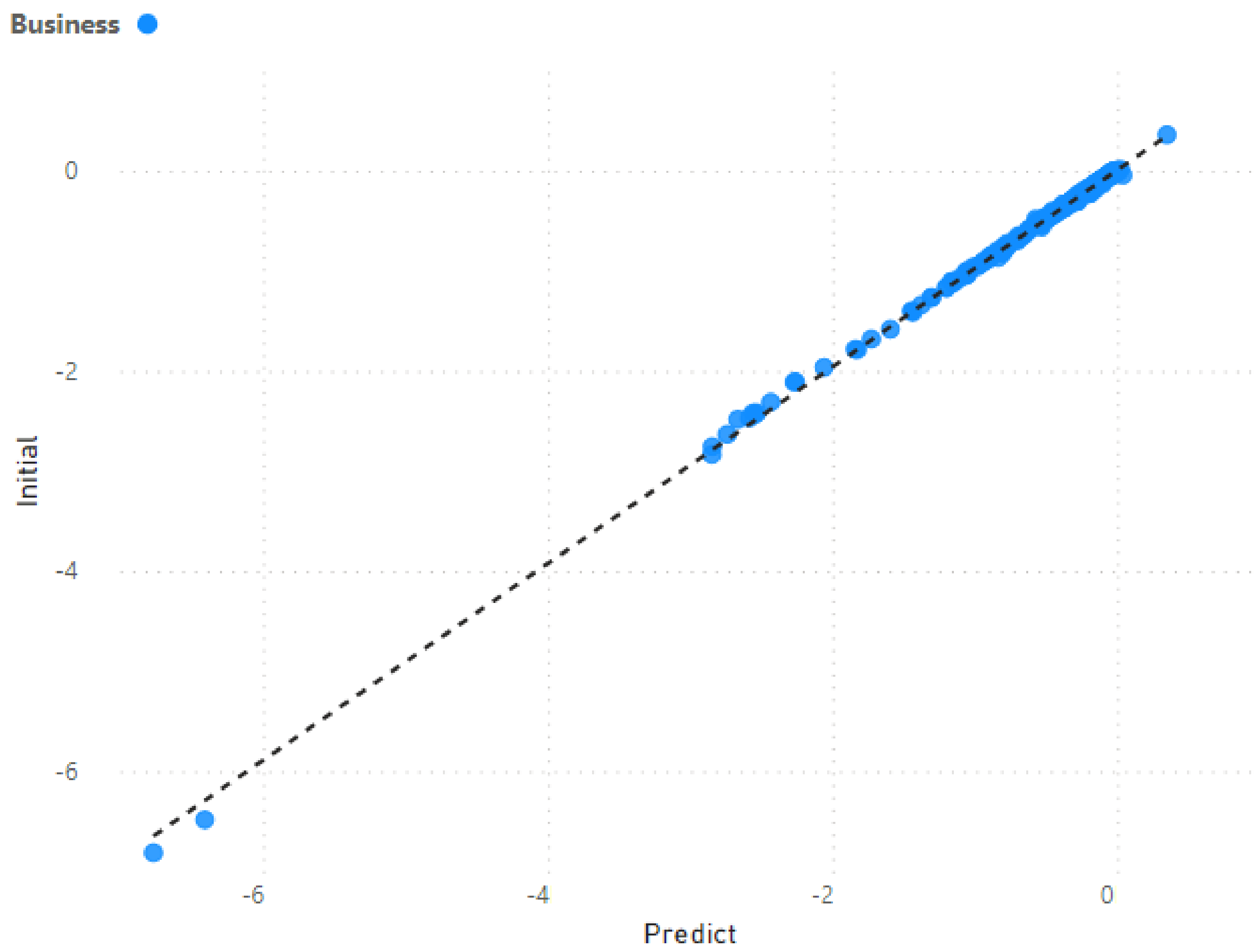


Figure 24 Scatter plot for the Pre-fitted Soft Voting Regression's forecasted and initial cost values.

After this, a second scatter plot, was created on overall level in order to visualise the differences between actual costs and forecaster costs. This was done to see if how the actual values are performing with the live dataset. This scatter plot in Figure 25, is a bit more scattered as there are some circles out of the line, but still in overall seems to be in linear line.

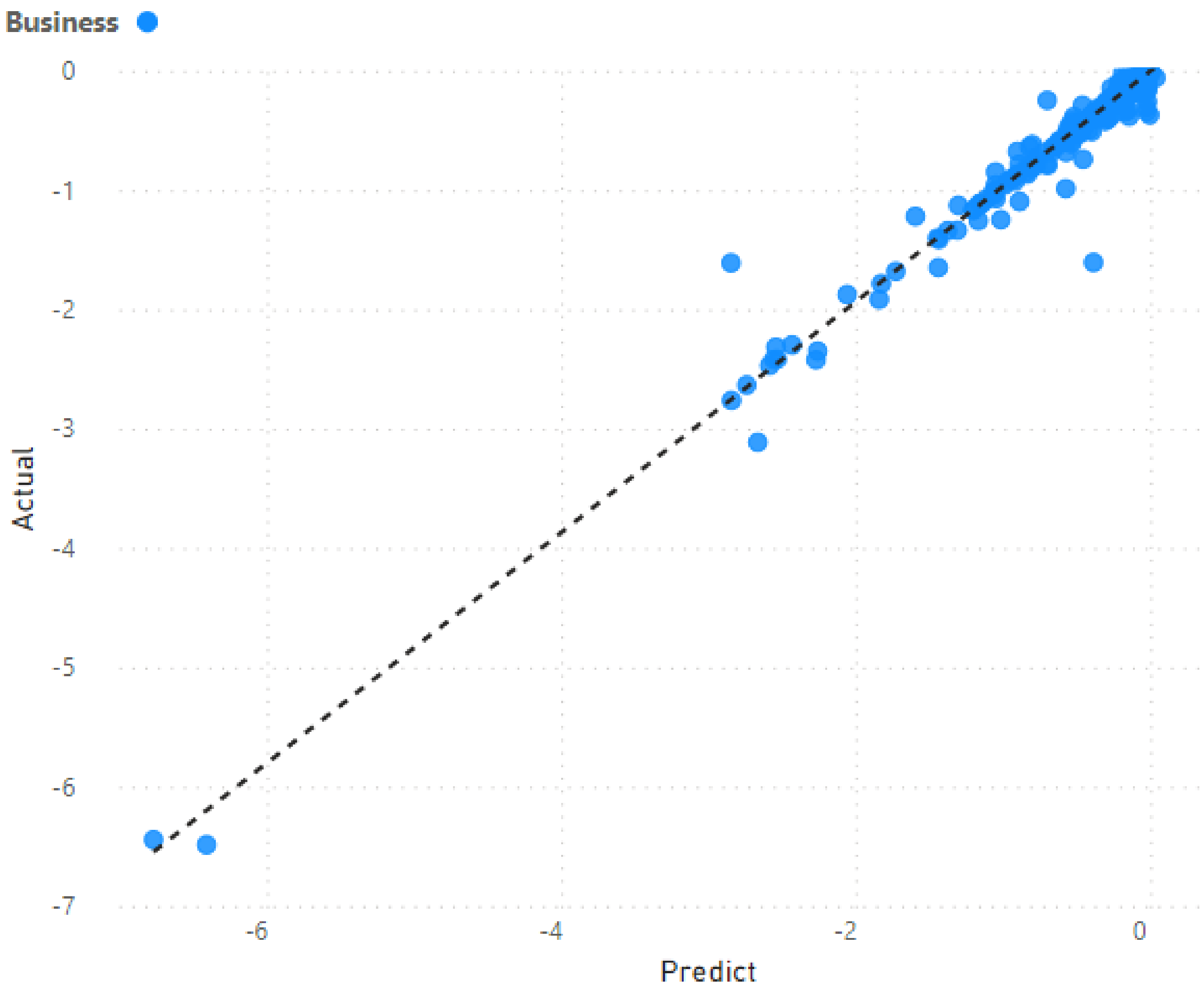


Figure 25 Scatter plot for the Pre-fitted Soft Voting Regression's forecasted and actual cost values.

Finally, as the difference between initial and forecasted value was almost linear, a scatter plot was created to visualize the difference between differences of initial and actual cost plan values, and difference between forecasted and actual cost plan values.

Scatter plot in Figure 26 was created in order to visualise how the differences between the real values and forecasted differences are performing. The results are very similar as in Figure 25. Naturally, this value should come to zero difference as much as possible, and many of the circles seem to be close to zero difference value.

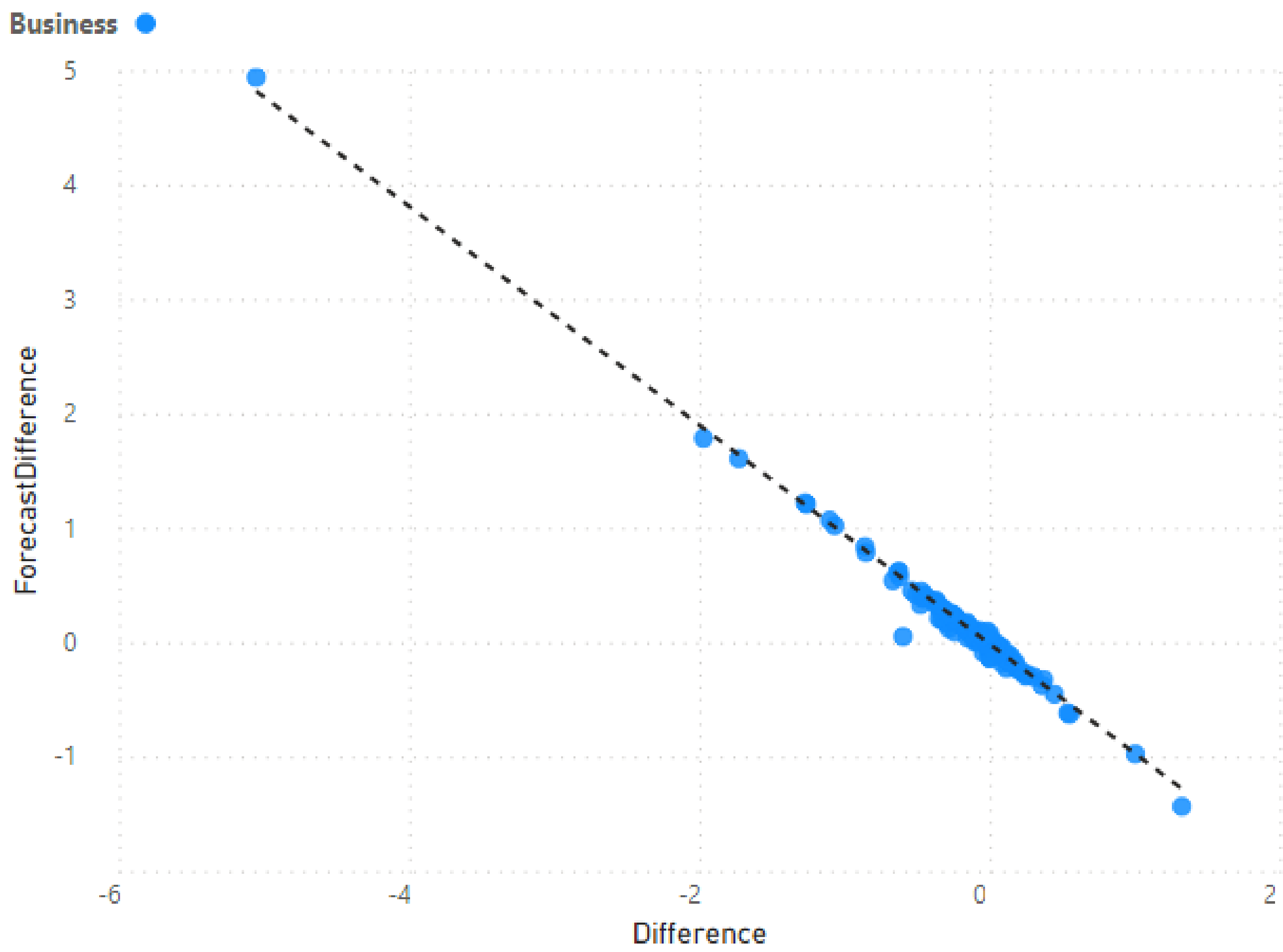


Figure 26 Scatter plot for difference between initial & actual cost and difference of forecasted and actual cost values.

4.3.2 Forecasting with Risky Project Characteristics

After the Pre-fitted Soft Voting Regressor model was confirmed to be close to linear regression, the risky project characteristics identified in chapter 4.2, were taken for analysis. The risky project characteristics were identified by where the most of the negative cost overruns are.

These confirmed risky project characteristics were identified as segments 9, 5, 6 and 7, whereas product type 12 is gather most negative cost overreuns, especially when delivering to country name 11. Risky project managers were identified as project managers 1, 2, 3, 4 and 5. All of the identified risky project charasteristics are listed in Table 2.

Name	**Risky Project Characteristics**
Segment 9	Yes
Segment 5	Yes
Segment 6	Yes
Segment 7	Yes
Product Type 12	Yes
Product Type 92	Yes
Product Type 95	Yes
Product Type 96	Yes
Product Type 97	Yes
Product Type 98	Yes
Country Name 1	Yes
Country Name 11	Yes
Country Name 2	Yes

Country Name 64	Yes
Country Name 75	Yes
Country Name 99	Yes
Project Manager 1	Yes
Project Manager 2	Yes
Project Manager 3	Yes
Project Manager 4	Yes
Project Manager 5	Yes

Table 2 Identified as risky project characteristics.

To capture the overall view on the risky characteristics scatter plots were created. These scatter plots show in a high level how the forecasting has performed when risky project characteristics have been selected gives some insight how this kind of forecasting could be used to predict possible cost overruns when risky project characteristics are identified.

Potentially underestimations shown in the circles below the trend line could cause major cost overruns in a project while overestimations could mean overbudgeting of the project.

By looking at the Figure 27, majority of the circles are in the linear line and some circles from segments 9 and 5 are a bit off, but overall, it can be said that the forecast fits quite well the actual costs.

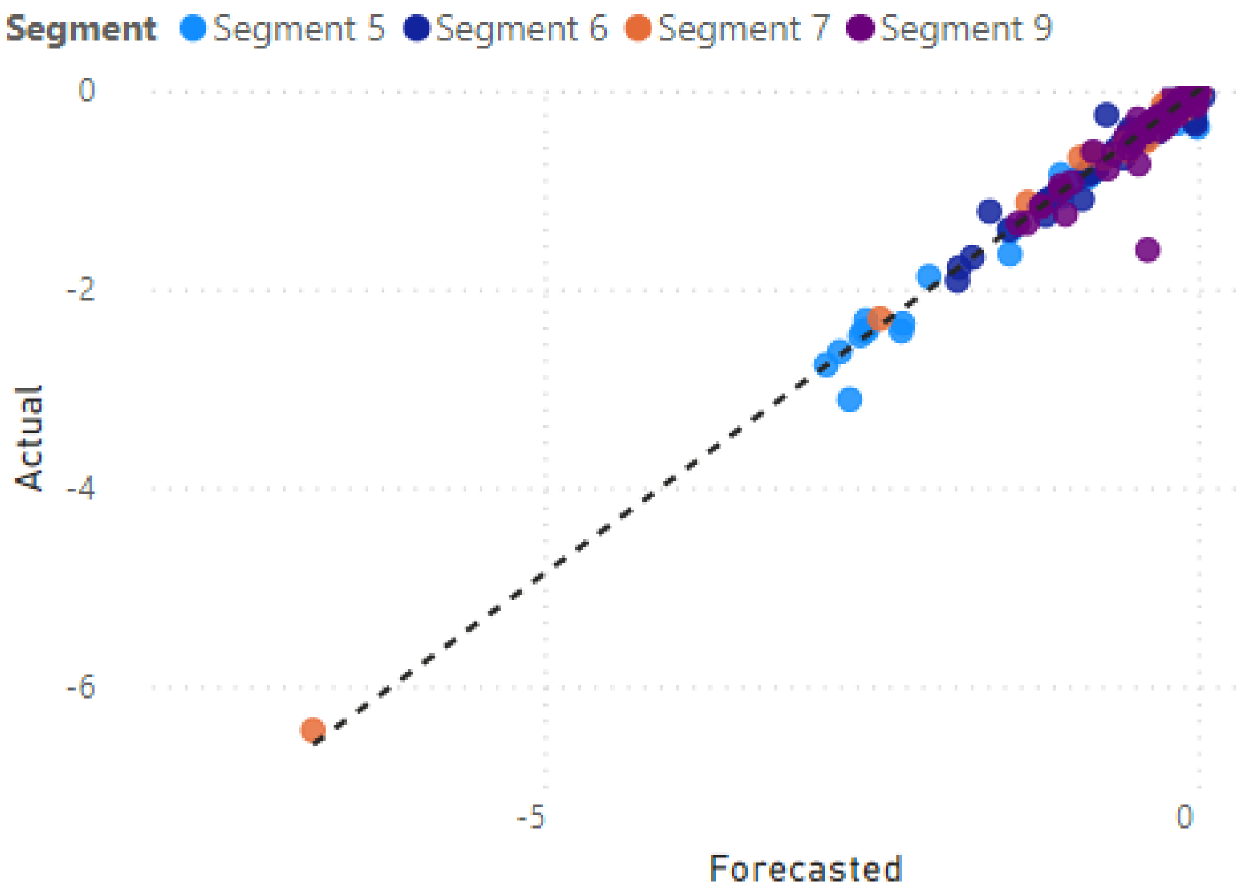


Figure 27 Scatter plot for difference between forecasted and actual cost values in risky segments.

Scatter plot in Figure 28, implicates that the overall the forecasted values are in line with the actual costs, but for especially risky country name 11 it seems that the forecasted values was underestimated in some circles.

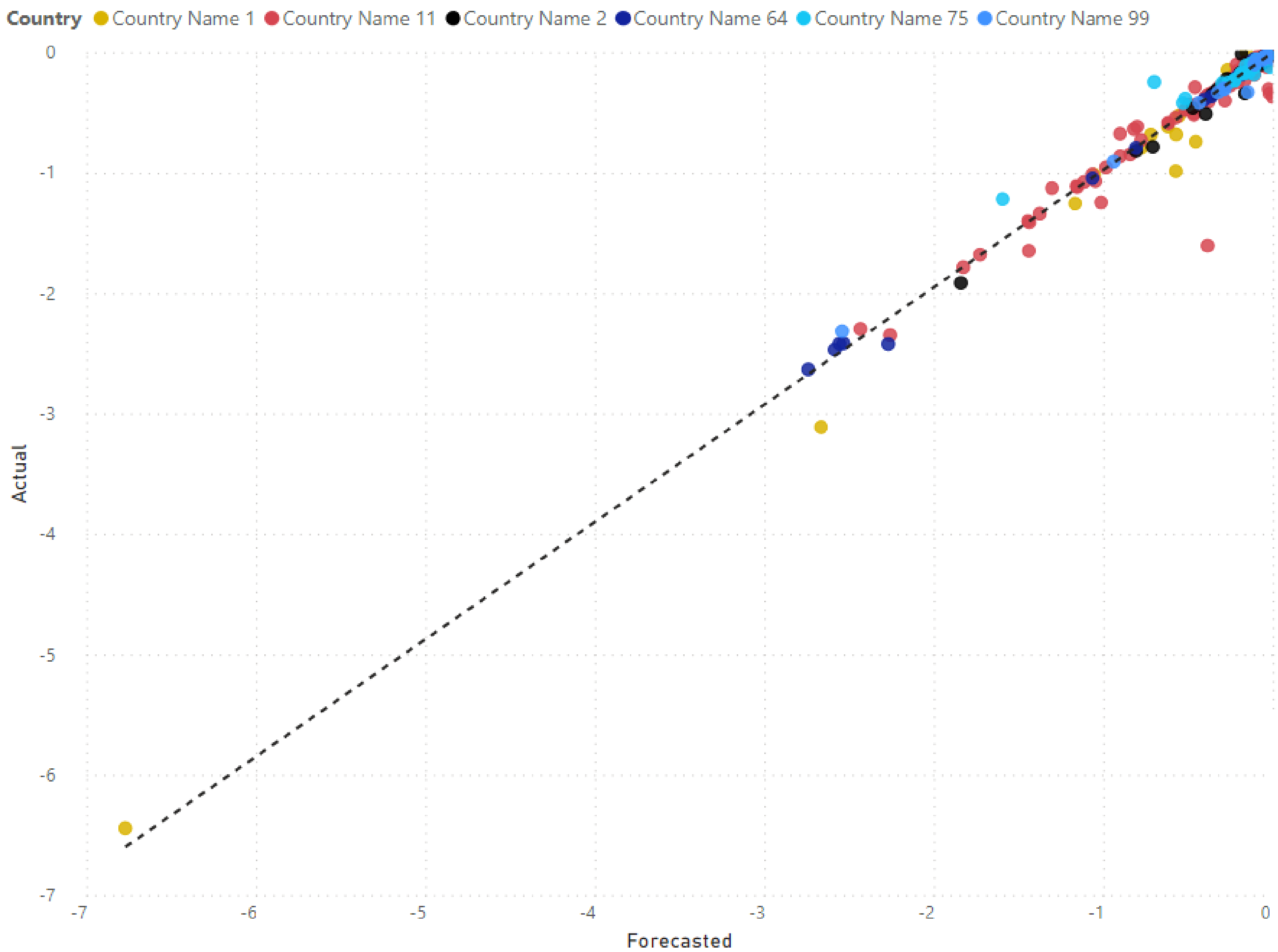


Figure 28 Scatter plot for difference between forecasted and actual cost values in risky countries.

Figure 29 explains how the riskiest products are performing in the forecast. Product 12, which was the riskiest product, seems to be difficult product to forecast. There are over-estimations and underestimations of the cost plan. This gives some implication, that the product might not be the only risky part of the corresponding project.

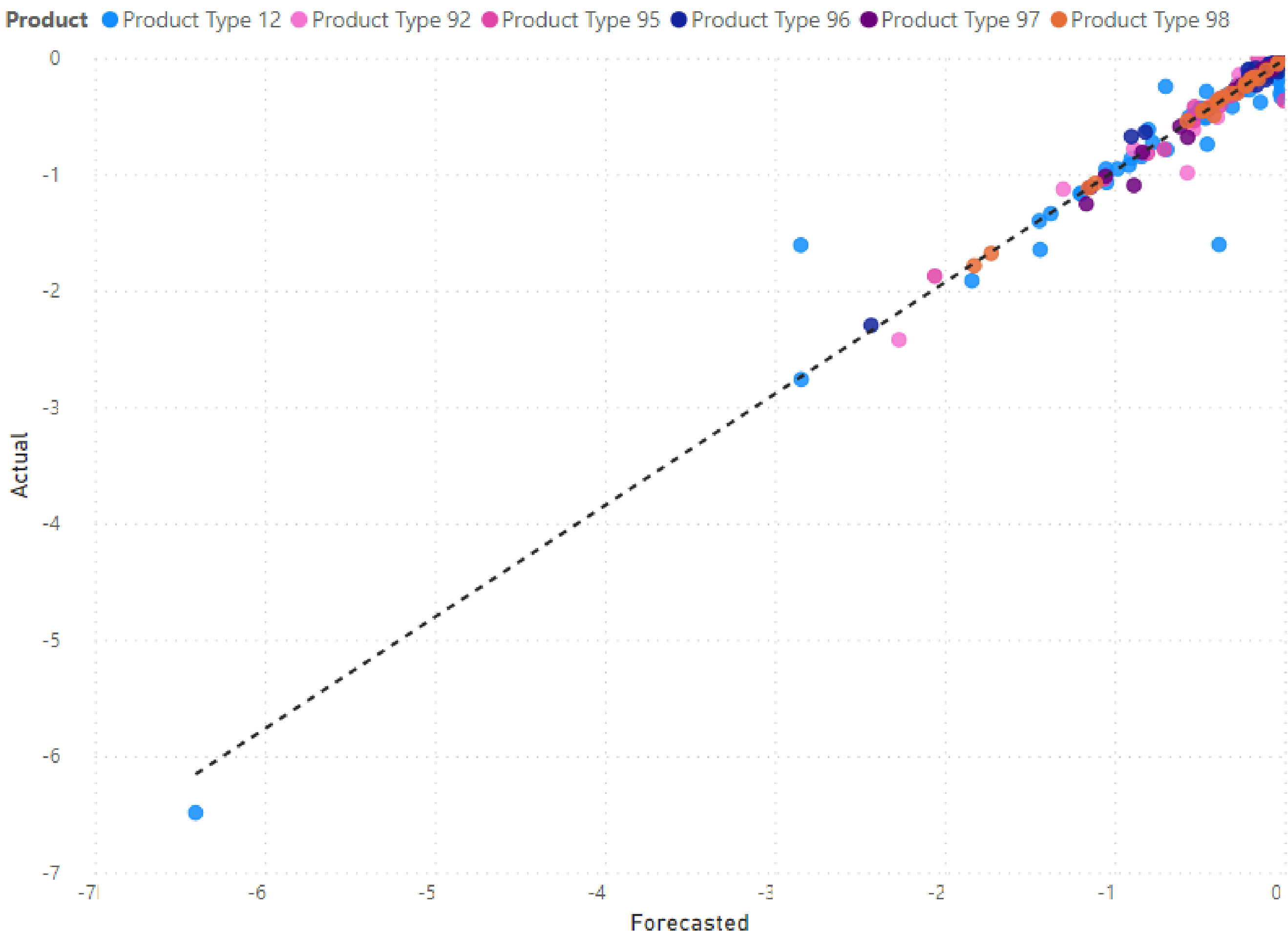


Figure 29 Scatter plot for difference between forecasted and actual cost values in risky products.

Scatter plot was created in Figure 30 also for the responsible project managers who had the most cost overruns to get insight on the effect of the nominated project manager on project costs. For project managers 1, 2, 4 & 5 are some larger underestimations by the forecasted cost plan, but also some overestimations are done for all project managers.

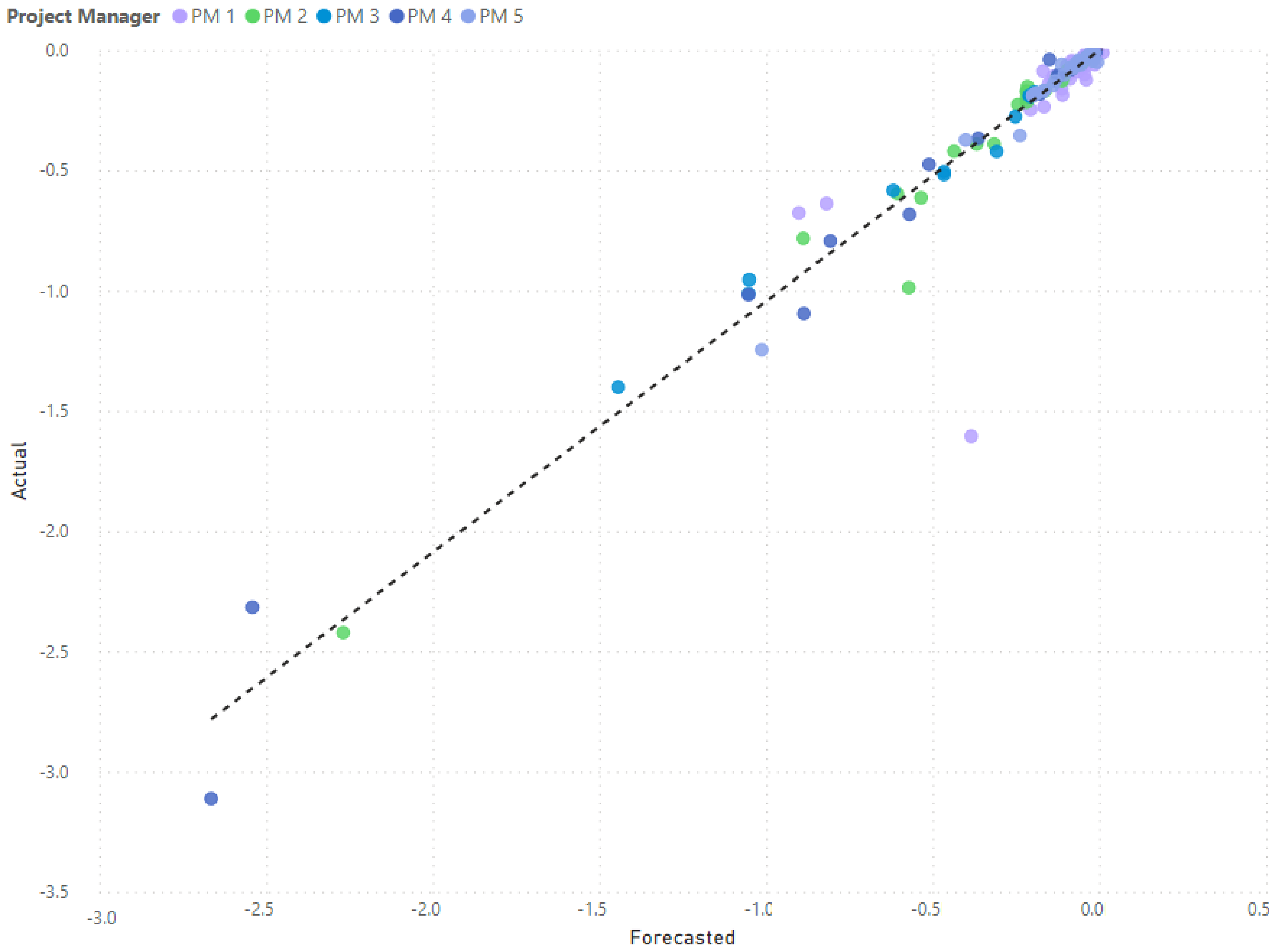


Figure 30 Scatter plot for difference between forecasted and actual cost values in risky project managers.

Taking a closer look at the forecasted values from the Pre-fitted Soft Voting Regressor model and compared to the installation segments identified as the risky segments in Figure 31, it can be noted down that the forecasted value is also quite precise on this level. For most of the risky segments, it seems to be almost the same value.

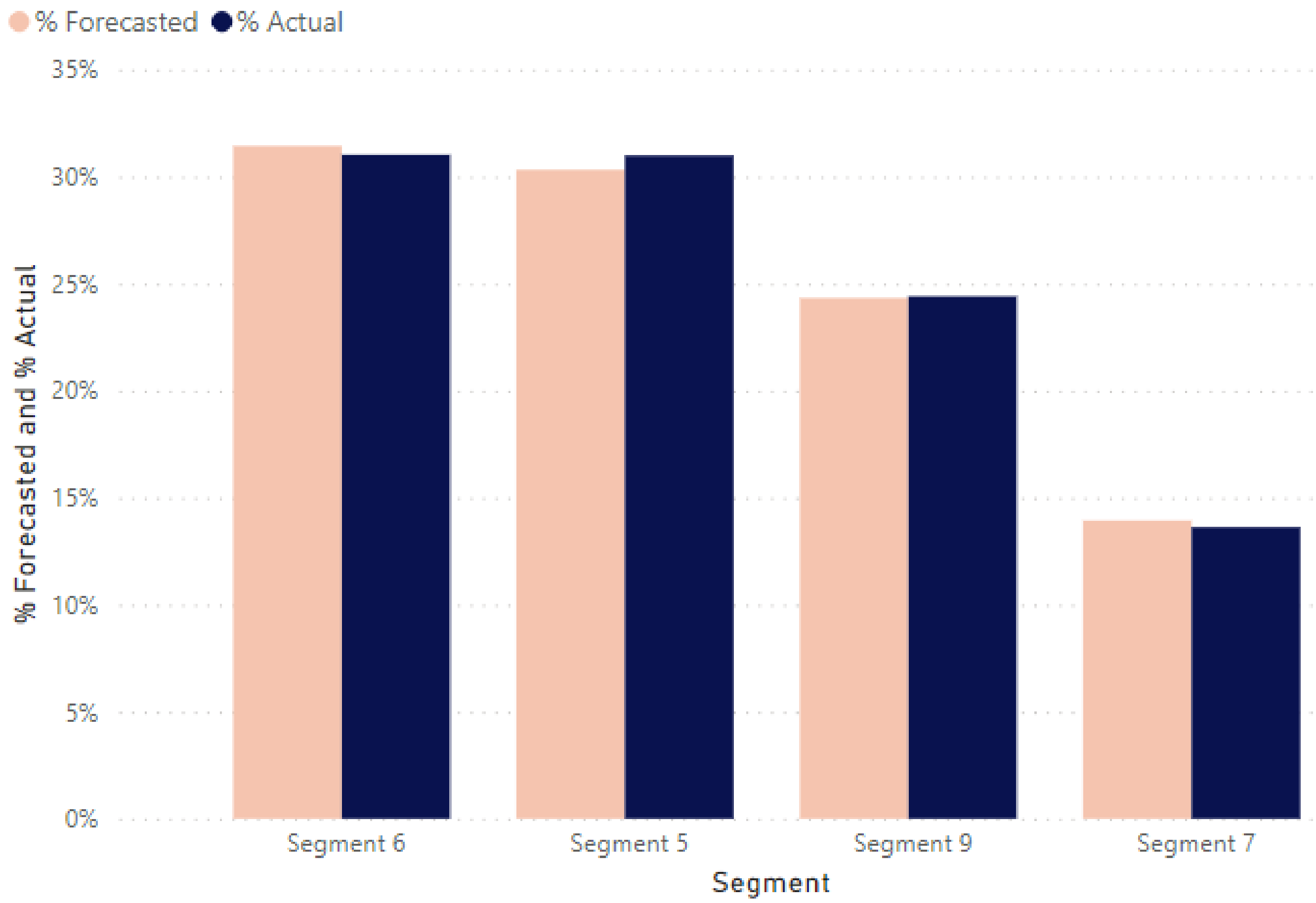


Figure 31 Differences in forecasted value and actual cost plan values per risky segments.

Comparing these values between forecast & actual cost plan values per segments, it seems that the forecast a bit higher on segments 6 & 7 whereas a bit lower than actual on segments 5 & 9. The forecast seems to be almost the same as actual on the segment 9, which was also the segment with the most total costs. Therefore, this bar chart gives some confidence that the results are reasonable.

By taking a look at Figure 32, all forecasts on the risky segments seem to be quite close to actual values. Underestimates are done by the forecast in segments 5 and 9. These could potentially then show less costs reserved in forecast than actual costs were. Segments 6 & 7 however show small overestimates which could then lead to less costs than forecast.

Segment	% Forecasted	% Actual
Segment 5	30.31%	30.98%
Segment 6	31.43%	31.01%
Segment 7	13.94%	13.62%
Segment 9	24.32%	24.40%
Total	**100.00%**	**100.00%**

Figure 32 Forecasted and actual cost plan values for risky segments.

To dig in deeper to the segment 9 in Figure 33, the vessel type 1 seems to be collecting the most actual costs. The forecasted cost plan value for vessel type 1 forecasts also a bit higher costs than actuals. The differences in all vessel types seem to be very small in overall level.

Installation Type	% Forecasted	% Actual Cost
Vessel Type 1	27.20%	27.11%
Vessel Type 9	24.58%	26.65%
Vessel Type 6	15.61%	15.26%
Vessel Type 2	9.20%	9.50%
Vessel Type 7	6.25%	6.08%
Vessel Type 5	4.84%	4.18%
Vessel Type 17	3.20%	3.60%
Vessel Type 15	2.76%	2.13%
Vessel Type 8	1.83%	1.50%
Vessel Type 11	1.59%	1.48%
Vessel Type 3	1.12%	1.04%
Vessel Type 13	0.87%	0.66%
Vessel Type 16	0.40%	0.43%
Vessel Type 4	0.33%	0.24%
Vessel Type 14	0.17%	0.11%
Vessel Type 12	0.06%	0.03%
Total	**100.00%**	**100.00%**

Figure 33 Vessel types for Segment 9 with forecasted values.

The segment 5, had the largest difference between the forecasted and actual cost plan values. The vessel types belonging to segment 5, have much more differences in the forecasted values than in segment 9 and these differences went up to 0,7%.

Installation Type	% Forecasted	% Actual Cost
Vessel Type 99	14.63%	15.04%
Vessel Type 98	15.40%	14.70%
Vessel Type 10	11.53%	12.18%
Vessel Type 97	11.80%	11.52%
Vessel Type 96	11.00%	10.94%
Vessel Type 95	7.03%	7.55%
Vessel Type 94	7.47%	7.22%
Vessel Type 93	7.47%	6.88%
Vessel Type 92	5.20%	6.06%
Vessel Type 91	2.12%	2.41%
Vessel Type 90	2.05%	1.65%
Vessel Type 89	1.55%	1.56%
Vessel Type 88	1.24%	0.95%
Vessel Type 87	0.63%	0.60%
Vessel Type 86	0.40%	0.33%
Vessel Type 85	0.34%	0.30%
Vessel Type 84	0.10%	0.05%
Vessel Type 83	0.03%	0.05%
Total	**100.00%**	**100.00%**

Figure 34 Forecasted and actual costs by vessel types in Segment 5.

The cost plan was forecasted also for the most common products delivered to risky segments 9, 5, 6 & 7. The costs of the product type 12 were a bit underestimated (-1,42%) compared to the actual costs. This could mean that costs in the segment are not purely caused by the product type 12.

Product Type	% Forecasted	% Actual
Product Type 12	19.40%	20.82%
Product Type 99	7.98%	7.72%
Product Type 98	7.02%	6.98%
Product Type 97	5.34%	5.59%
Product Type 96	5.14%	4.78%
Product Type 95	4.64%	4.76%
Product Type 94	3.80%	3.89%
Product Type 92	3.30%	3.36%

Figure 35 Most common products delivered to risky segments.

In Figure 36, the Country name 11 was forecasted a bit lower cost as well like product type 12 in comparison to the actual cost plan values in the risky segments and most common products. Overall differences in the country forecasts were relatively small, but as the portion of the projects in the country rise, the cost differences rise as well.

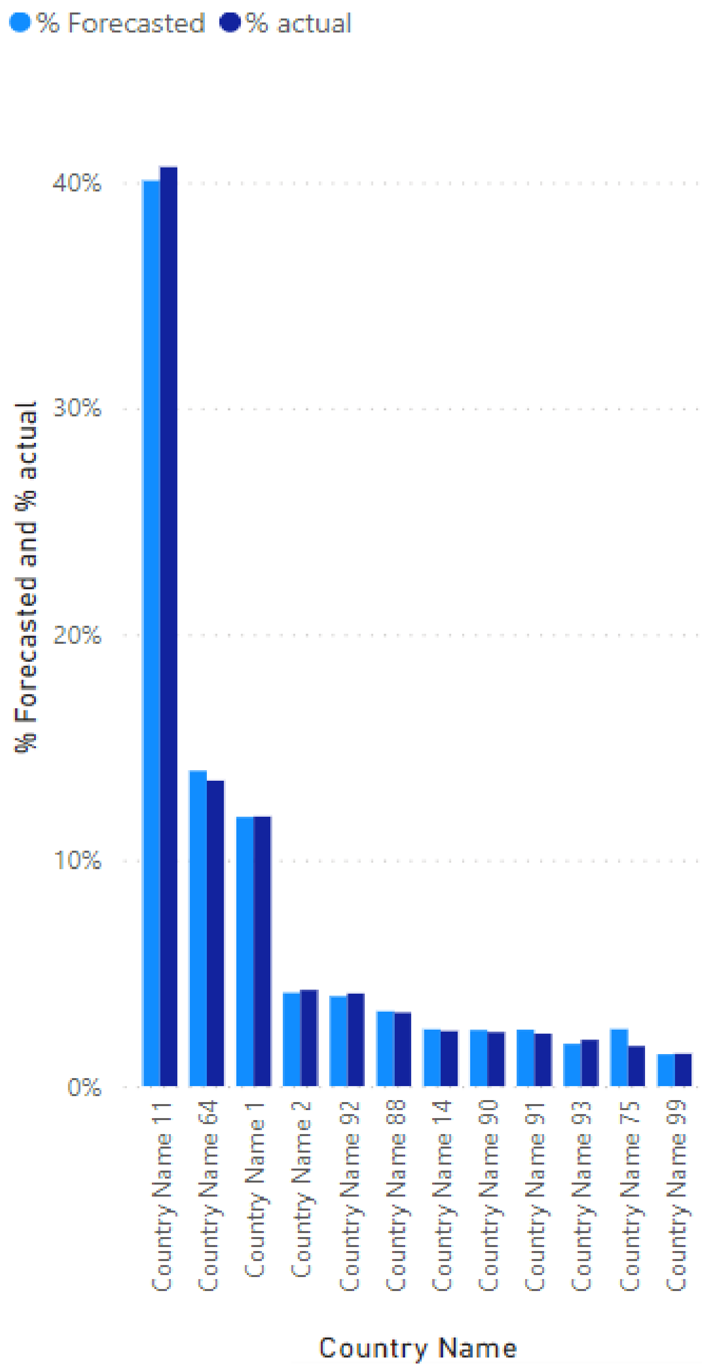


Figure 36 Forecasted and actual values per most common countries.

A scatter plot was created in Figure 37 to sum up previous results of risky characteristics in the projects in installation segments 5, 6, 7 & 9, delivered to customer in country name 1, 11, 2, 64, 75 and 99 with product type 12, 92, 95, 96, 97 & 98. These projects were delivered by project managers 1, 2, 3, 4 & 5. All of these characteristics listed in Table 1, have shown cost overruns which could be identified as risky characteristics.

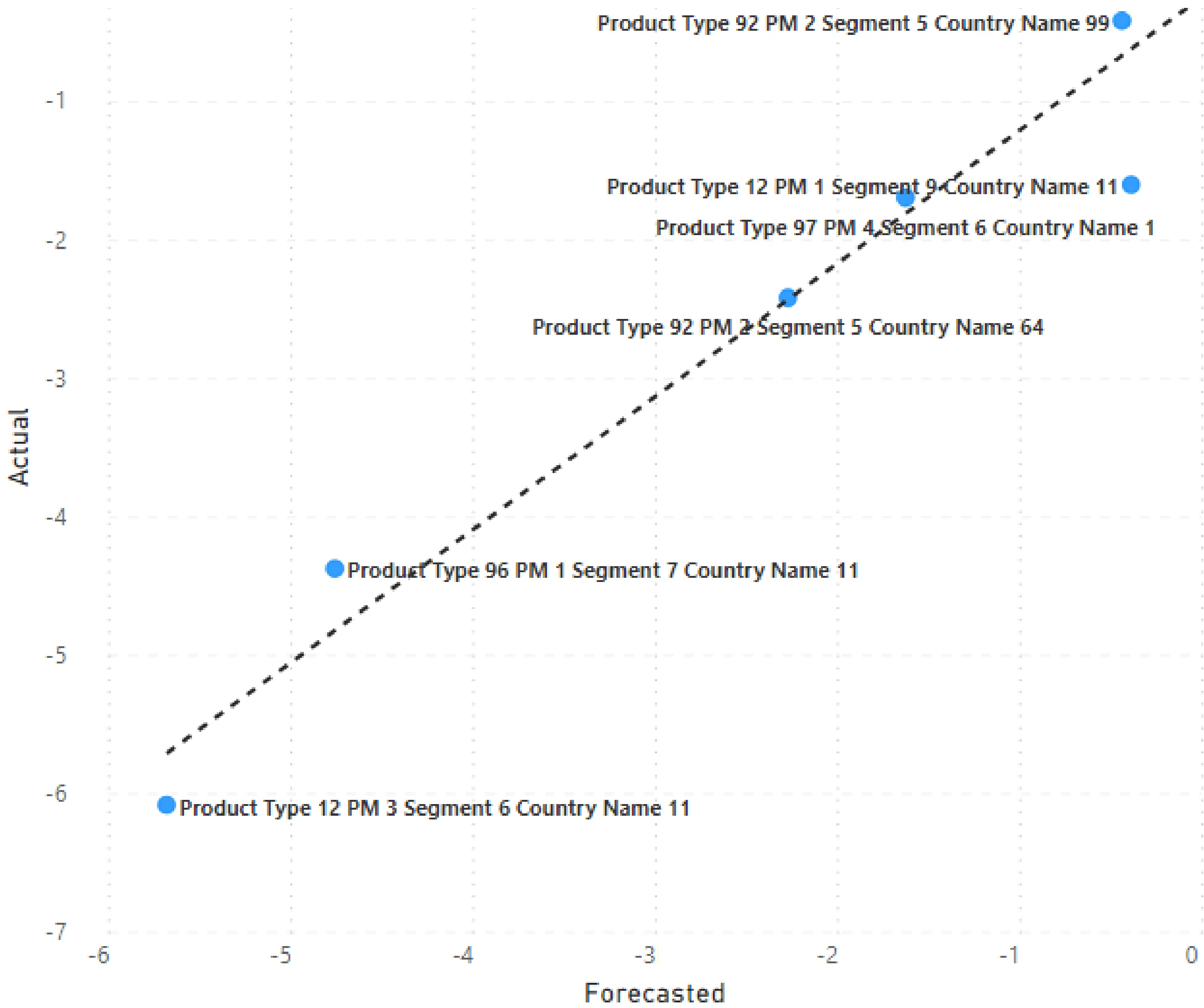


Figure 37 Scatter plot for difference between forecasted and actual cost values with all risky project characteristics.

Product Type	PM	Segment	Country Name	Forecasted	Actual	Difference
Product Type 12	PM 1	Segment 9	Country Name 11	2,52%	9,60%	-6,98%
Product Type 12	PM 3	Segment 6	Country Name 11	37,13%	36,38%	0,75%
Product Type 92	PM 2	Segment 5	Country Name 64	14,84%	14,48%	0,36%
Product Type 92	PM 2	Segment 5	Country Name 99	2,85%	2,51%	0,34%
Product Type 96	PM 1	Segment 7	Country Name 11	31,10%	26,17%	4,93%
Product Type 97	PM 2	Segment 6	Country Name 1	0,52%	0,44%	0,08%
Product Type 97	PM 4	Segment 6	Country Name 1	10,62%	10,15%	0,47%

Table 3 Differences between forecasted and actual cost values with all risky project characteristics.

By analysing Figure 37 and Table 3, it could be said that the forecast does fit for most of the cases, but there are certainly some circles which are not in line. Overestimations by the Pre-fitted Soft Voting Regressor model are naturally even welcome, but underestimations could potentially cause money loss.

As the Pre-fitted Soft Voting Regressor model implicated a 94% success result, this scatter plot seems to be near that result as well since the circles are very near to the trend line.

	Product	PM	Segment	Country	Forecasted
Product	1.0000000	0.1034936	-0.57992282	0.2627919	0.25838831
PM	0.1034936	1.0000000	-0.48412292	-0.1529946	-0.11241727
Segment	-0.5799228	-0.4841229	1.00000000	-0.5580371	0.06194379
Country	0.2627919	-0.1529946	-0.55803714	1.0000000	0.23910015
Forecasted	0.2583883	-0.1124173	0.06194379	0.2391002	1.00000000

Table 4 Correlation table of Risky Project Characteristics.

The correlation coefficient table for risky project characteristic values used in Figure 37 and Table 3 indicate that with these scenarios, there is some positive correlation between product and country, and product and project manager.

Significant negative correlations can be found on Product & Segment, Segment and Project Manager, Segment and Country. The correlation coefficient in a correlation plot visualises these results.

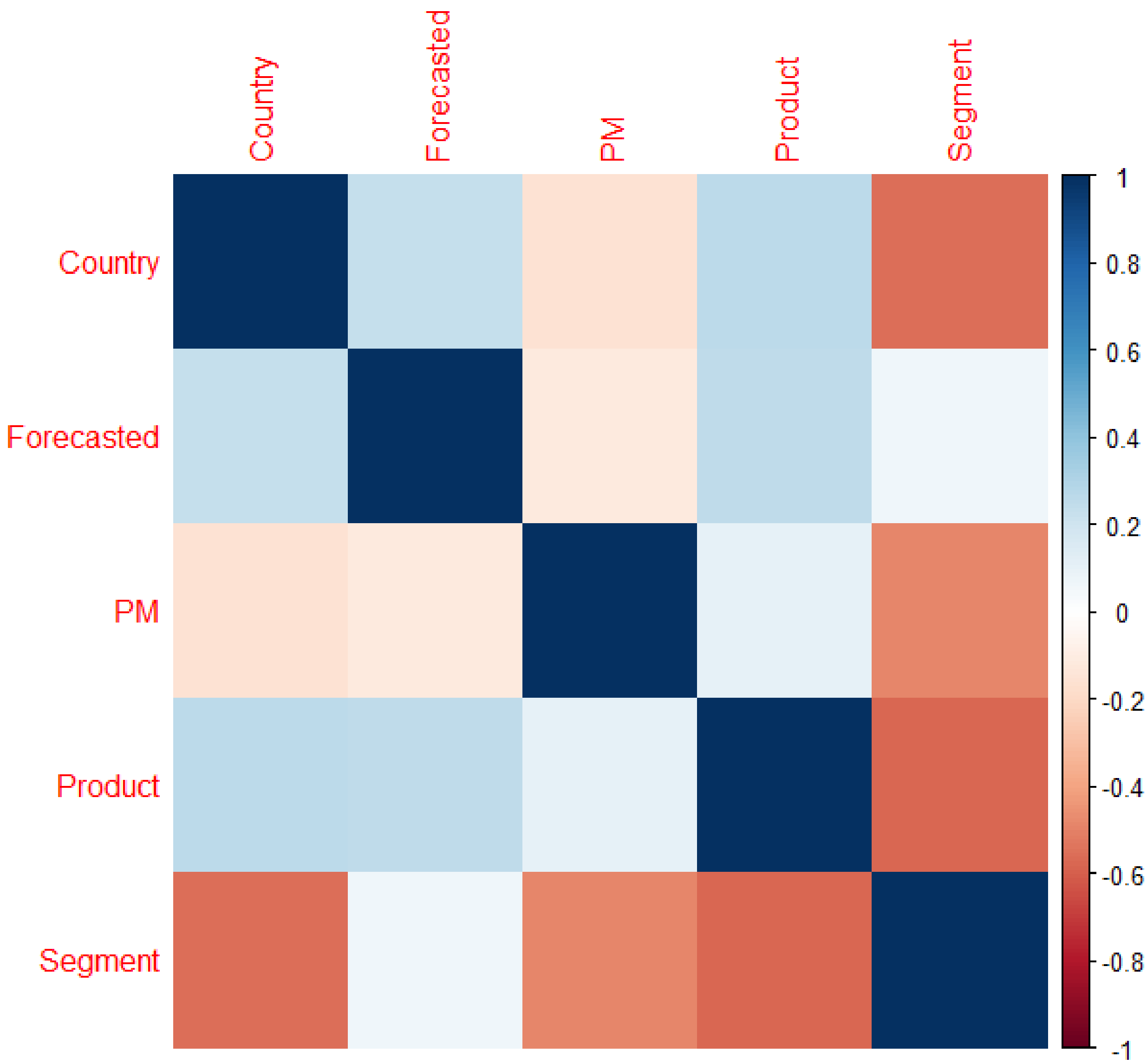


Figure 38 Correlation Matrix of Risky Project Characteristics.

5 Conclusions

This thesis researched the effects of positive and negative cost overruns in marine services industry projects for a global organization. The research focused on marine service projects on a global scale. The objective of the thesis was to find answers to questions especially such as how to identify risky projects characteristics and to provide this vital information to the organization for better cost planning of the projects. Previous research was used as a literature for the theoretical part and for the research the datasets were gathered from the organizations Enterprise Data Warehouse.

The research problem was to find out what is causing the differences between initial and completed project cost plans. The differences can be caused by difficulties in estimating the initial project cost plan budget in the first place, but also when there are risky project characteristics involved. These certain characteristics could jeopardize the initial cost plan and cause major positive and negative cost overruns.

Research questions initiated at the beginning of the research were such as:

Which type of project the marine service projects are?

What is the accuracy of the initial project cost plan estimates?

How to identify risky projects characteristics?

Marine service projects are projects where customer value is created by the project team working in co-operation with the customer to define problems and produce solutions to solve the problems. These are often unique type of problems to be solved as the service package is not often specified completely same. Technical expertise of the project team plays a major role in service projects.

The accuracy of the initial project cost plan estimates had some relationship of how much of the projects were in relationship of the characteristic. The more projects were in that characteristic, the greater the possibility was that there was a negative cost overrun. However, there were some characteristics which had less of the projects, but amount of negative cost overruns was larger.

The risky project characteristics were identified to certain segments, countries, vessel types, products, and project managers. These characteristics were identified by the objective linked to with most of the negative cost overruns. These areas where the negative cost overruns showed to some extend a pattern where a service work was done to a product in a certain customer country or vessel segment. The most important characteristic was the segment of the vessel.

The results of the Pre-fitted Soft Voting Regressor model forecast were promising. This kind of data-driven multidimensional model was the heart of the Pre-fitted Soft Voting Regressor model, and it could be used to similar forecasting purposes as to identify risky project characteristics and to forecast new values to those risky characteristics for better cost planning and preparation.

The results of correlation coefficient indicate that a there is in fact a relationship between the product for service work and the customer country to some degree. This could mean that some customer's from certain countries tend have certain products in their installations which may be causing risk to the projects. This could be also explained by the nominated project manager to the project. The combination of these risky characteristics is therefore something which should be kept an eye on.

This kind of research should be done more in the future as the research could be replicated also for other organizations where the project dimensional and financial data is available, and project cost plans are inserted. It is crucial for the organizations to identify these kinds of risky projects beforehand in order for better cost planning and preparation or even a declining of the customer contract.

5.1 Follow-up Research

This thesis' results provided information on estimation of project cost planning performance, and how to possibly identify risky project characteristics for any organization as the research is possible to replicate in any business. In case of global price inflation or war, force majeure situations for example, as the costs are rising heavily, the actual cost values might not be predictable anymore because the historical data is not reliable at that time.

As the dataset is available to the researcher, a regressor model can be easily created to replicate the forecast model. Correlation plot is recommended to find out the correlations between the risky characteristics. The forecast model can be then examined more closely on the identified risky characteristics for better planning.

The future follow-up research could focus on more on the specific risky characteristics and especially those connections or correlations between certain risky characteristics and how to use the information already before signing a contract where such risky characteristics are identified.

On a positive note, follow-up research could also investigate on positive cost overruns and how to plan better cost plan for the projects where positive cost overruns are likely. Better cost planning in non-risky project characteristics could also lead to money reservation savings in the organisation and free up resources.

References

Aramvareekul, P., & Seider, D. J. (2006). Cost-Time-Risk Diagram: Project Planning and Management. Cost Engineering, 48(11), 12–18. Available at: https://searchebs cohost-com.proxy.uwasa.fi/login.aspx?direct=true&db=buh&AN=23314446& site=ehost-live.

Artto, K., Kujala, L., Dietrich, P. & Martinsuo, M. (2008). What is project strategy? International Journal of Project Management. 26:1, pp. 4-12. Available at: https://doi.org/10.1016/j.ijproman.2007.07.006.

Ballard, C., Farrell, D., Gupta, A., Mazuela, C. & Vohnik, S. (2006). Dimensional Modeling: In a Business Intelligence Environment. International Technical Support Organization. International Business Machines Corporation.

Chang, A., Wiewiora, A. & Liu, Y. (2021). A socio-cognitive approach to leading a learning project team: A proposed model and scale development. International Journal of Project Management. 39:6, pp.646-657. Available at: https://doi.org/10.1016/j. ijproman.2021.05.003.

Chich, Y., Zwikael, O. & Restubog, S. (2019). Enhancing value co-creation in professional service projects: The roles of professionals, clients and their effective interactions. International Journal of Project Management. 37:5, pp. 599-615. Available at: https://doi.org/10.1016/j.ijproman.2019.04.001.

Dixit, V. & Tiwari, M, K. (2021). Learning curve based integrated procurement and project scheduling of multiple sister ships project. Computers & Industrial Engineering. 162. Available at: https://doi.org/10.1016/j.cie.2021.107691.

Dori, D. & Sharon, A. (2012). A Model-Based Approach for Planning Work Breakdown Structures of Complex Systems Projects. IFAC Proceedings Volumes. 45:6, 23-25.

pp. 1083-1088. Available at https://doi.org/10.3182/20120523-3-RO-2023.00255.

Dvir, D., Tzvi, R. & Shenhar, A. (2003). An empirical analysis of the relationship between project planning and project success. International Journal of Project Management. 21:2, pp. 89-95. Available at: https://doi.org/10.1016/S0263-7863(02)00012-1.

Fisher, O., Watson, N., Escrig, J., Witt, R., Porcu, L., Bacon, D., Rigley, M., Gomes, R. (2020). Considerations, challenges, and opportunities when developing data-driven models for process manufacturing systems. Computers & Chemical Engineering, (140). Available at: https://doi.org/10.1016/j.compchemeng.2020.106881.

Iranmanesh, S & Zarezadeh, M. (2008). World Academy of Science, Engineering and Technology. *Application of Artificial Neural Network to Forecast Actual Cost of a Project to Improve Earned Value Management System*. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.307.9199&rep=rep1&type=pdf.

Kwon, H. & Kang, W, C. (2018). Improving Project Cost Plan Estimation Accuracy and Preci sion by Analyzing Reserves for Both Identified and Unidentified Risks. Project Management Journal. Vol.50 (I) 86-100. Available at: https://journals.sagepub.com/doi/full/10.1177/8756972818810963

Lagemann, B., Lindstad, E., Fagerholt, K., Rialland, A. & Erikstad, S. (2022). Optimal ship lifetime fuel and power system selection. Transportation Research Part D: Transport and Environment. Vol 102. Available at: https://doi.org/10.1016/j.trd.2021.103145.

MARPOL, International Convention for the Prevention of Pollution from Ships (1973).

Available at:https://wwwcdn.imo.org/localresources/en/KnowledgeCentre/ConferencesMeetings/Documents/MARPOL%201973%20-%20Final%20Act%20and%20Convention.pdf.

Picciotto, R. (2020). Towards a 'New Project Management' movement? An international development perspective. International Journal of Project Management, 38(8), pp 474-485. Available at: https://doi.org/10.1016/j.ijproman.2019.08.002.

Review of Maritime Transport (2018). Available at: https://unctad.org/system/files/official-document/rmt2018_en.pdf.

Review of Maritime Transport (2021). Available at: https://unctad.org/system/files/official-document/rmt2021_en_0.pdf.

Siami-Irdemoosa, E., Dindarloo, S. & Sharifzadeh, M. (2015). Work breakdown structure (WBS) development for underground construction. Automation in Construction. 58. pp. 85-94. Available at: https://doi.org/10.1016/j.autcon.2015.07.016.

SOLAS, International Convention for the Safety of Life at Sea (1974). Available at: https://treaties.un.org/doc/Publication/UNTS/Volume%201184/volume-1184-I-18961-English.pdf.

Soler, E., Trujillo, J., Fernández-Medina, E., Piattini, M. (2008). Building a secure star schema in data warehouses by an extension of the relational package from CWM. Computer Standards & Interfaces, (30), pp. 341-350. Available at: https://doi.org/10.1016/j.csi.2008.03.002.

Svejvig, P. and Andersen, P. (2015). Rethinking project management: A structured litera ture review with a critical look at the brave new world. International Journal of Project Management, 33(2), pp. 278-290. Available at: https://doi.org/10.1016/j.ijproman.2014.06.004.

Tvedten, I, O & Baer, S. (2022). Retrofitting towards a greener marine shipping future: Reassembling ship fuels and liquefied natural gas in Norway. Energy Research & Social Science. 86. Available at: https://doi.org/10.1016/j.erss.2021.102423.