Tampere University

Niloufar Valinejad

MAINTENANCE COST ESTIMATION FOR LOAD HANDLING EQUIPMENT

Maintenance cost estimation for reachstackers using gradient boost regressor estimator

Master level
Signal processing
Joni Kamarainen
Juho Vihonen
Sep 20

ABSTRACT

Niloufar Valinejad: Prediction of cumulative maintenance costs for load handling equipment
Tampere University
Degree Programme
Sep 2020

Virtually all heavy-duty machines need maintenance, and the maintenance cost can be significant for providers of cargo and load handling solutions. Professionals in such industries must be able to forecast this cost accurately as minimizing this cost helps assure reasonable profits for companies and improving their economy.

The cost for maintenance varies for different machines based on the payloads they carry, working environment, operator skills, among others. That makes the cost difficult to predict since we usually do not have access to such sources of information. The current practice of using statistical regression methods cannot suitably capture the relationship between the repair cost of heavy equipment and its influencing factors.

In this thesis, the potential of Machine Learning (ML) models was evaluated as an alternative method for the prediction of maintenance cost of load handling machines. The distinctive difference is discovering the possibility of predicting this cost based on telemetry data merged with machine details, also analyzing parameters affects this cost the most.

This study was conducted based on data received from 483 Kalmar's reachstacker's since 2014 during their service contract or warranty contract. First, a detailed analysis of the historical data allows identifying the distributions of maintenance expenses and their fluctuated patterns during different RS' life periods. Then the research continued by the implementation of a tree-based ML model to predict two different predictive variables; 1) Cumulative maintenance cost per engine working hour (CMCPH) and 2) Cumulative maintenance cost per lift (CMCPL).

The results of the ML approach show better interpretability and adequate accuracy by considering CMCPL as the output variable with Meter per lift, fuel used per lift, and tons per lift as the most influential predictors of Maintenance Cost. One surprising observation was having the length of the service work order as one of the topmost important features affecting the result of the experiment. An accurate prediction of future equipment maintenance costs can promote decision-making tasks related to equipment budget and resource planning by injecting more observations to the model to decrease the variance.

Keywords: Gradient boosting regression, Maintenance cost, Heavy-duty machines, Data analysis, machine learning.

PREFACE

I would first like to thank my inspiring thesis advisor and examiner Associate Prof. Joni-Kristian Kamarainen from the Computing Sciences department at Tampere University for all of his supports and beliefs in me. I appreciate the trust that he put in me.

I would like to extend my deepest thanks to my caring and patient supervisor Docent Dr. Juho Vihonen from the Data-Driven Services team at Cargotec for his excellent technical guidance and feedback throughout this project. He always made me feel confident in my abilities and he was kindly encouraging me through this journey.

I would like to express my great thanks to dear Mr. Pekka Mikkola, the leader of the Data-Driven Services team, CIO Office, at Cargotec, who provided me the great opportunity of working with Cargotec professionals and creating the environment that I can learn, grow and flourish.

I would also like to thank my other colleagues at Cargotec for their wonderful cooperation. It was always helpful to bat ideas about my research around with them. I also benefitted from debating issues with my friends and family. If I ever lost interest, they kept me motivated.

Tampere, 8 September 2020

Niloufar Valinejad

CONTENTS

1. INTRODUCTION ..... **ERROR! BOOKMARK NOT DEFINED.**
1.1 Concepts and definitions 2
1.1.1 Cargotec 2
1.1.2 Reachstacker (RS) 2
1.1.3 Service work order (SWO) 3
1.1.4 Operating cost 4
1.1.5 Telemetry 4
1.2 Purpose 5
1.3 Scope of the study 5
1.4 Thesis organization 6
2. LITERATURE REVIEW 7
2.1 Literature search methodology and results 7
2.2 Cumulative cost model (CCM) 7
2.3 Life-to-Date repair cost solution 8
2.4 Period-Cost-Based solution 11
2.5 General regression neural network (GRNN) solution 13
2.5.1 General regression neural network overview 13
2.5.2 Modeling of equipment maintenance cost with GRNN 14
2.6 Summary 15
3. METHODOLOGY 17
3.1 Machine learning (ML) 17
3.2 Linear regression (LR) 17
3.3 Regression trees 19
3.4 Gradient boosting regressor (GBR) 19
3.5 Model Validation methods 21
3.5.1 Cross-validation (CV) 22
3.6 Hyperparameter optimization methods 23
3.7 Coefficient of determination (R Squared) 25
3.8 Practical implementation 25
4. EXPERIMENTS ..... **ERROR! BOOKMARK NOT DEFINED.**
4.1 Preliminary analysis 27
4.1.1 Data quality assessment 28
4.1.2 Feature exploring 29
4.2 Data Division 37
4.3 Linear regression 37
4.4 Gradient boosting regression 38
5. CONCLUSION 43
5.1 Recommendations 43
5.2 Final conclusions 44

REFERENCES 45

LIST OF FIGURES AND TABLES

Figure 1. Kalmar product portfolio development (Kalmar, 2019) 2
Figure 2. Super Gloria Reachstacker (Kalmar, 2019)..... 3
Figure 3. Service work order lifecycle 4
Figure 4. Geometric representation of the equipment's CCM by Vorster..... 8
Figure 5. Life-to-Date methodology concept by Mitchell, 1998..... 10
Figure 6. Period-Cost-Based methodology concept (Mitchell, 2011)..... 12
Figure 7. Illustration of 5-fold cross-validation..... 23
Figure 8. Grid search visual representation 24
Figure 9. Cargotec Conceptual AI/ML Architecture..... 26
Figure 10. SWO cumulative cost distribution for all available RS's in every 8 months of their lifetime..... 27
Figure 9. Abnormal drops on cumulative attributes of three reachstackers. 28
Figure 10. CMC per number of engine hours for eight RS's..... 29
Figure 11. Implementation of Vorster model for all RS's based on engine working hours..... 30
Figure 12. Grows in maintenance cost of RS's with shutdown engines..... 31
Figure 13. RS's that have been working for a long time without needs for expensive maintenance 31
Figure 14. Correlation between the number of lifts and engine working hours ... 32
Figure 15. CMC per number of lifts for all RS's..... 33
Table 1. Influential variables on CMC 33
Table 2. Variables of interest retrieved from Table 1 based on engine working hours. 34
Figure 16. Correlation of Engine working hours with cumulative attributes (Kilometre, Tons, fuel consumption)..... 34
Table 3. Variables of interest calculated based on the number of lifts..... 35
Table 4. Attributes in final dataset derived based on the number of lifts.... 35
Table 5. Attributes in final dataset derived based on engine working hours..... 36
Figure 19. The probability distribution function of 2 datasets..... 37
Figure 17. Predicted values versus the observed values for CMCPHR by the LR model on the left and residuals vs predicted values of CMCPHR on the right..... 38
Figure 18. Predicted values versus the observed values for CMCPL by the LR model on the left and residuals vs predicted values of CMCPL on the right..... 38
Table 6. List of tunned hyperparameters and the range of values..... 39
Table 7. Best values for hyperparameters in each trained model..... 40
Table 8. Summary of $r2$ scores resulted from the GBR model..... 40
Figure 19. Predicted values vs actual values in CMCPL test set..... 41
Figure 20. Predicted values vs actual values in CMCPHR test set..... 41
Figure 21. The relative importance of features..... 42
Figure 22. Learning curve on training set calculated based on lifts 43

LIST OF SYMBOLS AND ABBREVIATIONS

TCO	Total Cost of Ownership
EMC	Equipment Maintenance Costs
CHE	Cargo Handling Equipment
RS	Reachstacker
SWO	Service Work Order
ERP	Enterprise Resource Planning
CCM	Cumulative Cost Model
LTD	Life-To-Date
PCB	Period-Cost-Based
CMC	Cumulative Maintenance Cost
GRNN	General Regression Neural Network
ML	Machine Learning
LR	Linear Regression
GBR	Gradient boosting regressor
CV	Cross-Validation
PDF	Probability Density Function
CMCPH	Cumulative maintenance cost per engine working hour
CMCPL	Cumulative maintenance cost per lift

1. INTRODUCTION

Load handling equipment provides the functions of transporting all types of cargoes in different ports and container terminals worldwide. From the moment a fleet starts running, it is usually under massive workloads which causes the necessity of lifelong repair and maintenance. It is the same as other types of vehicles to stay in reasonable running conditions since equipment unavailability causes huge expenses for equipment owners.

On the other hand, the cost of maintenance is one of the most considerable expenses of the machine's Total Cost of Ownership (TCO). It makes customers cautious not only about the purchase price but also about all the future costs that equipment will cause them, such as labor, spare parts, repair, fuel consumption, tires. Therefore, load handling equipment providers stay competitive in the market based on the type of offered after-sales services.

An accurate prediction of equipment maintenance costs (EMC) paves the way for budget planning for equipment repair and spare parts. Better predictability of maintenance costs would benefit maintenance providers to stay cost-effective and determine suitable maintenance strategies on different occasions and give them the ability to offer more reliable service contracts, which increase customer satisfaction.

Predicting EMC remains challenging since it can significantly change depending on the type of payloads they carry, working environment, operator skills, equipment' age, reliability level, and other influencing factors in machines. We could have had an accurate estimation if we had access to all these attributes for machines with similar working conditions. However, not all the relevant information like the operator skills or working environment is available. As digitalization gains ground, the vast quantities of data from multiple sources are accessible: telemetry data received from machines, service work details with labor cost, spare part cost. The massive flow of data will offer new ways to predict the future of a fleet maintenance cost. Taking advantage of gathered data from various equipment and resources and implementing data mining models provides a more reliable approach to predict EMC and generating more significant cost-efficient maintenance systems by identifying potential high-cost equipment.

1.1 Concepts and definitions

In terms of having a better understanding of the terminology of the research, some of the definitions and concepts are explained below.

1.1.1 Cargotec

Cargotec Oyj is a Finnish company founded in 2005. It produces cargo-handling machinery for ports, terminals, ships, roads, and local distributions. It has three different business units: Kone Corporation's container handling (Kalmar Global), load handling (HIAB), and marine cargo handling (MacGregor). The equipment we analyze in this research is manufactured by Kalmar business unit which is a provider of cargo handling equipment and automated terminal solutions, software, and support services (Kalmar - Cargotec, 2019).

1.1.2 Reachstacker (RS)

A reachstacker is a versatile piece of Cargo Handling Equipment (CHE) used for handling intermodal cargo containers in terminals or ports. Reachstackers (RS) are popular in container terminals and ports because of their flexibility and great stacking capacity. They are available in different types to transport all sorts of containers, flat racks, and sling loads up to 45 tonnes (Josse, 2017) and piles them quickly and efficiently in various rows depending on its access. The capacity and technology of different RS's classes are shown in Figure 1.

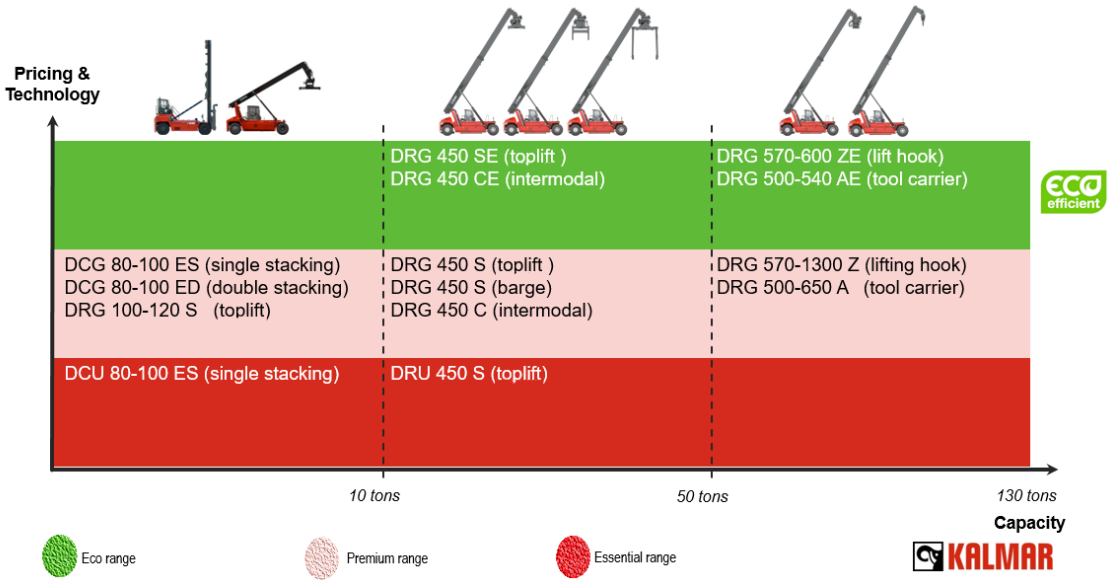

Figure 1. *Kalmar product portfolio development (Kalmar, 2019)*

Figure 2 is a model of RS named Super Gloria. In this research, we gathered data from RS's equipped with telemetry devices (see section 1.1.5) regardless of their models or capacity.

Figure 2. *Super Gloria Reachstacker (Kalmar, 2019)*

1.1.3 Service work order (SWO)

A Maintenance or Service Work Order (SWO) is a lifecycle of a demanded maintenance task in four main phases: identification, creation, completion, and recording as it is shown in Figure 3.

These phases can be broken into smaller tasks and bring about a smooth maintenance process that ensures tasks do not get stuck in one state and turn into the backlog (Cousineau 2019). In this research, SWO data provides information about changed spare parts, duration of maintenance, start date, and finish date, whether the equipment was under a service contract or warranty contract. The cost of the maintenance, which contains labor price summed with spare parts price and some description about how the maintenance task progressed.

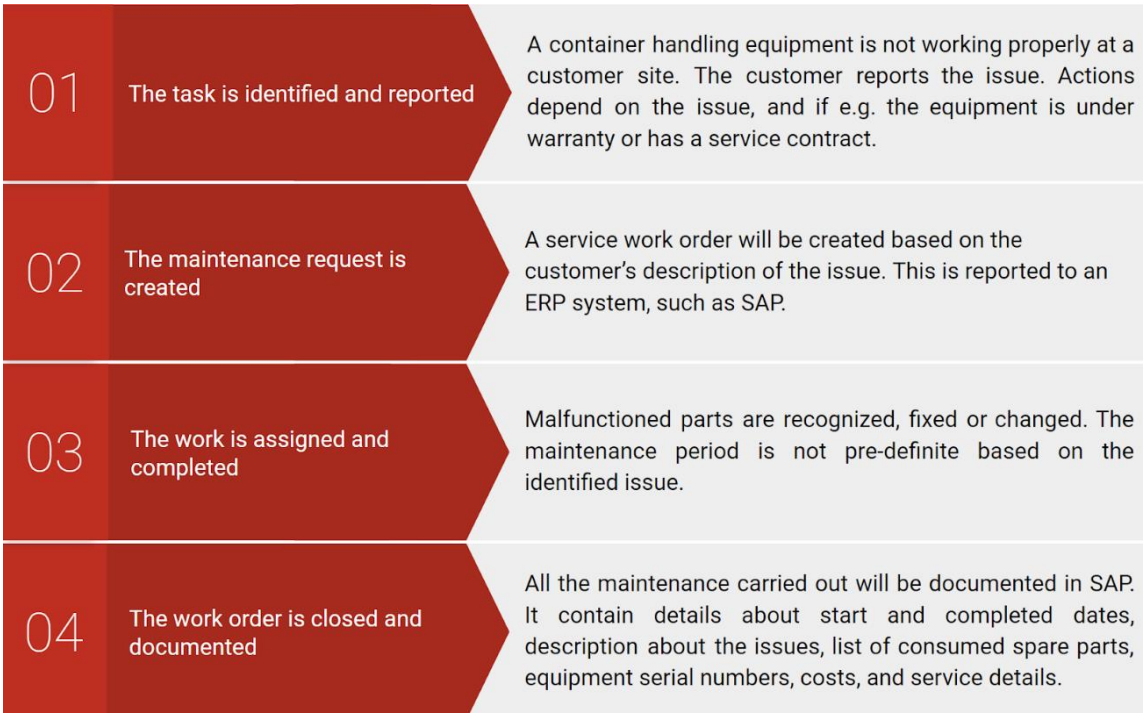

Figure 3. *Service work order lifecycle*

1.1.4 Operating cost

Operating cost is the sum of expenses that a machine cost for the owner while working on projects. These expenses are 1) Fuel, 2) Lubricants, filters, and grease, 3) Repairs, 4) Tires, 5) Replacement of high-wear items (Robert et al. 2018). A Kalmar machine is under warranty contract in its first two years of the working life, and it mostly covers the expenses of lubricants, filters, grease, and some part of repairs. However, various service contracts offer different types of service and maintenance opportunities to customers during and/or after the warranty contract.

The operating costs we will track in this research are the result of the equipment maintenance that Kalmar offers to its customers during warranty and service contracts which includes the cost of spare parts and labor.

1.1.5 Telemetry

Some remote sources are able to transmit data measured by sensors. This automatic measurement and wireless transmission of data are called telemetry. Depends on the type of the sensors, data, which might be voltage, temperature, pressure, and such, are combined as signals and sent to a remote receiver. Upon reception, the signal is converted to the original elements, and the user processes them as their interest (Rouse, 2005).

Nowadays, heavy machinery such as RS's is equipped with telematics devices that collect measurable data from sensors within machines and send them remotely. This data is called telemetry data and contains engine working hours, different parts of the machine temperature, liters of fuel consumed, kilometers that machine moved, the number of lifts, the number of tones was lifted, among others.

1.1.6 Enterprise resource planning (ERP)

Enterprise resource planning (ERP) is the integrated management of the main working areas of an organization's business processes by a centralized system. ERP gives the possibility of using a central database for software components available in different modules with fundamental business areas, such as HR, production, finance, marketing, maintenance, management, supply chain management, and customer relationship management. Choosing core modules depends on the companies preferences and related to their particular business.

Modules in the organization access to the information which is shared by other modules. Therefore, companies that using ERP are saved from data redundancy entries, and it will bring about collaboration and accuracy into different departments (Rouse, 2019). SAP (Systems, Applications, and Products in Data Processing) is an ERP software that is used in Cargotec. This system is made by SAP SE, a multinational software corporation that is a market leader in the field of ERP solutions (Rouse, 2019).

1.2 Purpose

This study aims to explore and analyze how maintenance costs are affected by data gathered from RS's configs and their workload properties. In other words, we want to minimize the maintenance cost of the machine by keeping eyes on the maintenance costs of the early life stages of the machine. In order to do that, we must detect the most significant parameters in potential high-cost cases with the help of machine learning techniques.

1.3 Scope of the study

The study is limited to evaluating data from Kalmar's RS telemetry data, maintenance details available in the Cargotec SAP system, and vehicle configuration data. The collected data is limited to 483 vehicles equipped with a telemetry device and their service work orders recorded from the year 2014 to 2019. It is noticeable that the result of the

research is applicable to other types of heavy machinery with telemetry devices such as forklifts, terminal trucks, and more of the same.

The study does not attempt to predict the future maintenance cost of the machines. Rather it focuses on early warnings of potentially high-cost cases to be able to minimize their maintenance cost in the future. This study also tries to find the most influential factors on the maintenance cost.

1.4 Thesis organization

The structure of the thesis is further divided as follows. Chapter 2 presents a literature overview of different researches on the prediction of heavy machinery's maintenance costs. Chapter 3 explains the techniques and machine learning methods applied in this research. Chapter 4 explains the data and the relation between attributes with visualization, training the implemented model, and discussing the results and evaluation of the used methods in terms of parameters and performance in this research. This chapter also includes other aspects of the project, such as feature extraction. Finally, chapter 5 presents a summary of the research, conclusions, future perspectives, and further research in EMC prediction.

2. LITERATURE REVIEW

In the load handling machinery industry, the two most well-known issues that managers should deal with are costs and customer satisfaction. Moreover, it is noticeable that the cost of maintenance is usually the largest single element of machine cost. It makes up between 15% and 20% of the total equipment budget. The repair cost constitutes 37% of machine cost over its service life (Yip et al. 2014). Therefore, precise maintenance planning and estimation of its cost play critical roles as a management activity and provide companies profitable, stable business and reduce the overall cost of operating for their customers.

2.1 Literature search methodology and results

This research started by studying the documents and brochures of different pieces of equipment produced by Cargotec Oyj and its business areas Kalmar, Hiab, and MacGregor, to gain a better understanding of the problem, characteristics of machines and their use cases. The research continued by finding related papers from TUNI library databases. In order to find related papers, some keywords are used, such as maintenance cost prediction, load handling equipment economy, maintenance management, heavy equipment operating costs. Five papers are selected among all the academic literature based on their similarity to this research.

2.2 Cumulative cost model (CCM)

This model was proposed first by Vorster (1980). The CCM model is used to analyze some specific equipment management decisions such as the initial purchase decision, production capacity replacement, retire/replace decisions, maintenance strategy analysis, capital rebuild decisions, repair cost analysis. This model uses equipment age as the main means to calculate the average cost. The average cost is the sum of the cost of the equipment since it starts working to the point of the desired date. All are owning and operating expenses that a manager of the machine may consider as influential on the economic life of equipment can be under CCM, such as purchase price, fuel, repair, spare parts.

As shown in Figure 4, CCM usually creates a bathtub shape as a result of assigning machine age to abscissa and the average cost to the ordinate. We can see that Vorster

originated the average cost curve at the cumulative cost that represents the purchase price of the machine.

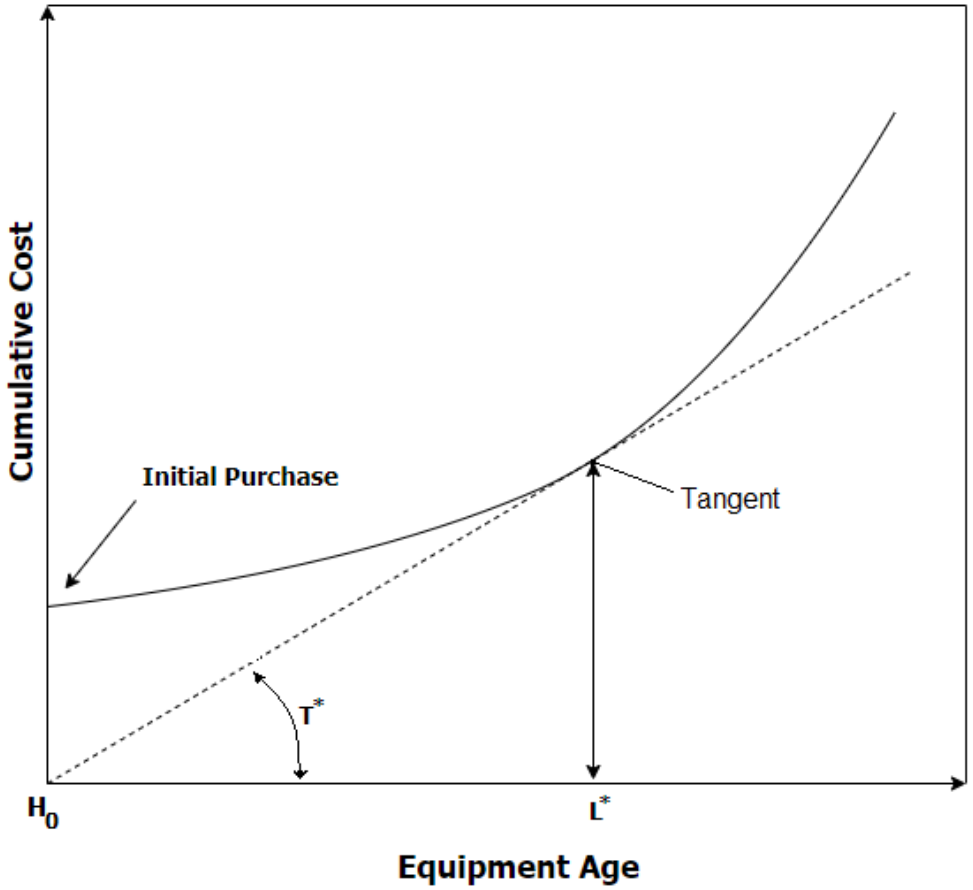

Figure 4. *Geometric representation of the equipment's CCM by Vorster*

Tracking the CCM is depicting the optimum economic life of the equipment. It is shown as L^* in Figure 4 and is defined by a geometric tangent to the curve, and it is drawn from the origin. T^*, is the lowest average cost for a fleet achieved when the slope of the tangent, L^*, is reached. In this model, age can take three forms of calendar age, age in cumulative hours of use, or age which is based on production units.

2.3 Life-to-Date repair cost solution

The life-to-date (LTD) solution is a regression model represented by Mitchell (1998) to solve the problem of heavy machines repair cost based on their age in cumulative hours of use. Mitchell used field data of 260 construction machines from 4 different companies to process the behavior of costs that could be applied throughout the industry. In this study, seventeen groups of equipment were made based on their size and type. He found that there is a second-order polynomial curve, shown in Equation 1, which is the

best fit for his field-collected data, although it did not perform as well for all equipment in a company neither within some groups of similar machines.

$$CC_{p\&l} = A * H_w + B * H_w^2, \qquad (1)$$

in this equation, $CC_{p\&l}$ is the cumulative cost of labor and spare parts for machines from zero working hours up to H_w hours. A coefficient is a linear portion that shows the growth of cumulative cost over time, and coefficient B is the inflection of the cumulative cost curve. For a given A coefficient, a smaller B coefficient means a greater economic life expectancy of the machine. Each data point must be a pair of meter/hour reading (H_w) and the cumulative maintenance cost (CMC) at the time. It is suggested to pick one data point for each machine in sufficiently large fleets to have better statistical results. In this case, it is better to have machines with different age ranges, so that the data points are spread equally throughout the expected economic lifespan of that fleet.

However, if the size of the fleet is not large enough, several data points can be picked for each machine to satisfy the economic lifespan by considering the following. Firstly, an equal number of data points are needed for each machine to ensure the same level of influence of each machine on the experiment. Moreover, picking data points spread evenly throughout the life of the equipment would be ideal (Mitchell, 1998).

By plotting $CC_{p\&l}$ as ordinate and H_w as abscissa, Mitchell fit a second-order polynomial curve and applied the intercept through the origin to perform a linear regression. Coefficients A and B can be obtained by Matrix multiplication or trend line function in Equation 1. Extraction of repair cost accumulation and totals are possible after building this model.

Figure 5 is a representation of the LTD methodology, and it shows that CMC's can be found with cumulative working hours H_w using Equation 1. By drawing a straight line from H_0 (new machine) to H_w, the average maintenance costs per working hour can be calculated for the whole period using Equation 2:

$$a = A + B * H_w. \qquad (2)$$

The marginal repair cost per working hour can be calculated at any cumulative working hours, Hx, by taking the derivative of Equation 1 with respect to H_w, and Equation 3 will be the result of it:

$$m = A + 2B * H_x, \qquad (3)$$

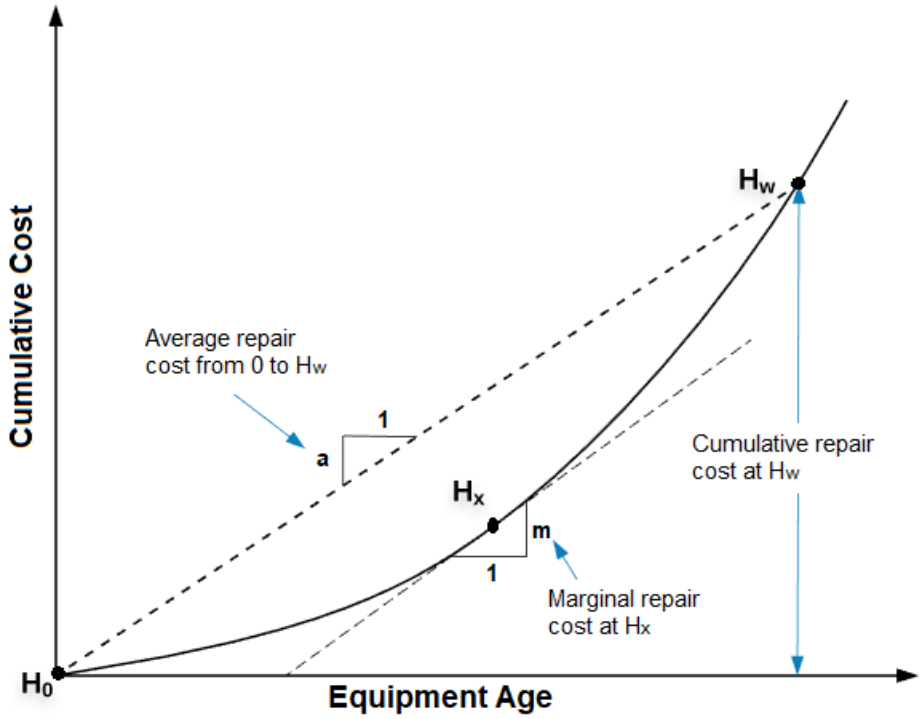

Figure 5. *Life-to-Date methodology concept by Mitchell, 1998*

where m denotes the slope of the line tangent to the curve. Using the LTD model to estimate equipment repair costs can be considered as a powerful technique. One of the main advantages of LTD is that it picks the various experiences of a single or group of machines. A collection of low, high, and average maintenance costs are captured throughout the life of the machine. It makes the data reliable and representative of what can be expected from similar equipment in its different life's moments. This model can be applied to individual machines or different equipment family.

Equipment managers can plan for the maintenance budget by applying the estimated number of working hours of a machine to the model in a given time. The LTD methodology can be considered as a data-driven repair forecasting tool that can be applied to a various range of machines to predict maintenance costs. However, it is more accurate when applied over the course of time for machines with a particular job. In order to find the repair cost, the user of the model first needs to declare coefficients A and B then equations are applicable to each equipment depends on their working hours.

In Mitchel's research, the average coefficient of determination, R^2, for each equipment category was about **0.72**, which can be considered too low. The other drawback of this model is that the repair data for equipment should be available from its initial purchase until the current time. In this model, variability in repair costs was considered more explicable as a function of hours of use and less affected by the type of use, while the impact of operating conditions is undeniable. Although the LTD solution applies to all

machine types in a company simultaneously, better predictive models will be the result of analyzing machines with almost the same size and type against one another. In case of a large sample size breaking down the fleets by equipment manufacturer offered in this research to have better accuracy. The LTD methodology may provide an accurate prediction of the maintenance cost if data points are evenly spread throughout the machine's age spectrum. If a large number of a specific class of machines in a company are purchased or replaced at the same time, the age spectrum available for analysis could be limited (Mitchell, Hildreth and Vorster, 2011).

2.4 Period-Cost-Based solution

The period-cost-based (PCB) solution is another methodology based on the relation between the cumulative cost of labor and spare parts with the cumulative number of machine working hours. Knowledge of the cost of repair parts and labor of a machine or class of machines for any period of machine lifespan should be available in order to implement this model. The period is defined by the amount of machine working hours from the start point at H_s until the end of the period, H_e.

Considering the curve drawn from the CMC and the number of machine working hours is a second-order polynomial, the PCB method is using the mean value theorem to estimate the cost of a machine in a specific timespan (Mitchell, Hildreth and Vorster, 2011).

Based on the mean value theorem for a function $f(x)$ that continues between two close, bounded endpoints $[H_s, H_e]$ and it is differentiable over the open interval (H_s, H_e), there is at least a point Hm in this interval at which tangent to $f(x)$ is parallel to the straight line passing through defined endpoints (Strang and Herman, 2016) such that:

$$f'(H_m) = [f(H_e) - f(H_s)]/(H_e - H_s)]. \quad (4)$$

As shown in Figure 6 for a second-order polynomial, H_m is almost located in the middle of H_s and He curve. To calculate the average maintenance cost between times H_s and H_e, availability of the maintenance cost details of the machine in this period, and the length of the period is mandatory. The average cost of repair parts and labor is shown in Figure 6 as slope m. As shown in the picture, m is also the slope of the line tangent to the curve. This line is the indication of the marginal maintenance cost of the machine at the time H_m. As mentioned earlier, Equation 1 is defining this curve. Taking the derivative of Equation 1 with respect to working hours resulted in Equation 3.

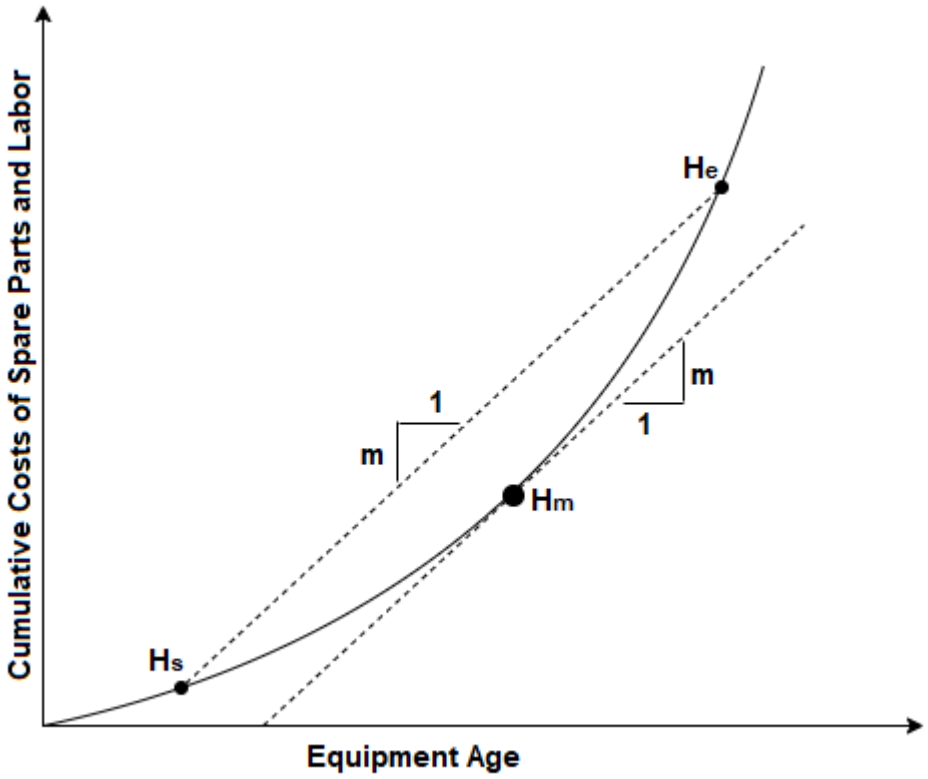

Figure 6. *Period-Cost-Based methodology concept (Mitchell, 2011)*

This differential defines the relation between marginal maintenance cost and machine working hours. The fact that the average maintenance cost from H_s to H_e is equal to the marginal cost at H_m makes the estimation of A and B coefficients possible by using the PCB solution. Each value of H_m with its corresponding maintenance cost value (not cumulative cost) can be indicated with linear regression techniques as $y = E + Wx$, and coefficients E and W can be calculated as $E = A$ and $W = 2B$.

The values calculated for A and B can be used to complete the equation available between CMC and machine working hours. The data needed to take advantage of PCB is mostly available within companies' database platforms, which makes PCB solution handy. Moreover, data collection for the PCB method can be started at any point in the machine lifetime. The possibility of analyzing the maintenance cost based on data gathered from the partial history of the machines increases the opportunity of data availability and expands the dataset. Older equipment which is more of the interest when it comes to estimation of maintenance cost is vastly available in the dataset, since removing incomplete data is not a necessity anymore.

However, the estimated cost between the period of H_s and H_e with the PCB method may not accurately represent the cost of the machine at that specific time. In this method,

fluctuations in the data have a direct effect on the result. For instance, expensive and uncommon maintenance costs that machines experience at a point in their lifelong will falsely increase estimation, while the performance of machines that experience severe maintenance before the period under study is excellent. Therefore, they cause an unusual decrease in cost estimation. As much as possible data should be used when developing the PCB model to estimate CMC to reduce bias in the result (Mitchell, Hildreth and Vorster, 2011).

2.5 General regression neural network (GRNN) solution

In this solution, the time-series approach, combined with a general regression neural network model, was applied to predict the maintenance cost of construction equipment. Time-series approaches take advantage of the fluctuation patterns and recent history of cost changes in the prediction models. A comparison of traditional linear and nonlinear GRNN models by Yip (2014) reveals that: multivariate time series modeling with fuel consumption can give a better description of the association between the current value of the maintenance cost and historical observations of both maintenance costs and related explanatory time series.

2.5.1 General regression neural network overview

GRNN's are memory-based single-pass neural networks proposed by Specht (1991) that estimate continuous variables. They use a Gaussian activation function in their hidden layer. They are used to suggest a nonlinear relationship between the target variable and a set of independent explanatory variables.

In this model, the output will be estimated based on the average of the output of the training data. The value of the target variable is calculated by taking the weighted average of the values of its neighbor's points. It makes the impact of close neighbors more than the distant ones. A chosen radial basis function (RBF) such as Gaussian distribution can be used to evaluate the level of neighbor's influence (weight) in which the input is the distance, and the output will be probability value, Weight = RBF(Distance). The standard deviation of this Gaussian distribution determines the influence of neighbor's points on the target variable. It is because of larger standard deviations makes the distribution curve more spread and the other way around. An optimization method can be used to reach an ideal standard deviation value.

The advantages of using GRNN models include its outlier handling capability and accuracy, even though modeling is based on a small dataset. This learning algorithm is used

to determine the relationships among data in time series, intervention variables, and relevant time series.

$$Y_t = f\begin{pmatrix} Y_{t-1} & Y_{t-2} & \dots & Y_{t-n} \\ X_{1(t-1)} & X_{1(t-2)} & \dots & X_{1(t-n_1)} \\ X_{2(t-1)} & X_{2(t-2)} & \dots & X_{2(t-n_2)} \\ \dots & \dots & \dots & \dots \end{pmatrix}, \quad (8)$$

where

Y_t Current observation

Y_{t-i} Previous n observation, $i = 1, 2, 3, \dots, n$

X_i Related time series or invention variable i.

$X_{i(t-j)}$ Historical observations of explanatory time series or invention variable at $(t-j)$.

n_i Correlated lagged values of related time series or invention variable i.

In equation 8, the value of the current observation is based on the value of its n previous observations and also the value of n last related historical variables, $X_{i(t-j)}$.

2.5.2 Modeling of equipment maintenance cost with GRNN

Yip (2014) collected a raw dataset of monthly total maintenance cost since 1998 for modeling. Fuel consumption as additional information on equipment operations was considered as influential factors on improving the accuracy of maintenance cost prediction. The amount of fuel consumption had a correlation with accumulated equipment operational duration and workloads. Changes in equipment fuel consumption usually cause changes in an equipment maintenance cost with or without lagged effects (Yip et al, 2014).

GRNN was used for multivariate and univariate models for the maintenance cost of construction equipment. Lag length optimization is determined by using the Akaike information criterion (AIC) to take into account the impact of historical observations. The maintenance cost series was divided into two parts of the validation dataset and training dataset. The twelve out-of-sample values, which represent the maintenance cost of the last 12 months, were used as the predictable period (validation dataset), and all earlier observations are used for training the model. The mean absolute percentage error (MAPE) over the actual and predicted values, was used for model accuracy evaluation:

$$MAPE = \frac{\sum_{t=N-M+1}^{N} \left| \frac{X_t - X_{t-1}}{X_t} \right|}{M} \quad (9)$$

where

N the number of observations in time series

M the number of test data

X_t the observed data at time t

X_{t-1} the predicted value of X_t based on the period of observation until X_{t-1}.

Lower MAPE value indicates smaller deviations between the forecasted and actual values of the time series. The predicted value of each out-of-sample prediction is used for the prediction of the next value in the one-step-ahead approach.

Overall, univariate and multivariate GRNN's predict two time series with decent levels of accuracy, with average MAPE of 24% and 20.4%, respectively. It is evident that multivariate GRNN performs slightly better accuracy with the input of historical maintenance cost in alongside the time series of fuel consumption. The reason behind it is that univariate GRNN uses a linear model to describe the serial relationship within a time series. In contrast, multivariate modeling used a nonlinear GRNN learning algorithm to depict the underlying complex relationship. However, GRNN can be trapped in the local minimum of the error surface and might stop training, although the global minimum of error has not yet been reached because of the iteration of searching an optimal smoothing parameter. This can be considered as a GRNN's drawback. Besides, neural networks are black-box models, and it reduces the capability of the analysis of the time-series dynamics. No consensus of method exists in determining the lag length for GRNN models in the time series approach.

2.6 Summary

Duo to the enormous impact of maintenance repair cost on equipment resource and budget planning, research investigation on this matter becomes wide. Surveying the literature on equipment maintenance costs shows that several approaches have been considered to predict the future expenses of the equipment. What these approaches have in common is the type of data they are using. The maintenance cost of machines in their lifetime to the date was the main attribute to estimate the future cost. Indeed, there are vast differences among various types of heavy machinery, their features, and the workload they carry, which may affect the expenses. Moreover, it is noticeable that they were not considered in the mentioned researches. It can be extracted from this chapter that there are still lots of gaps that can be the topic of research in this field of study.

As our research is based on a real case study in a heavy-load machinery supplier company, it is even more important to propose a methodology that can be applicable in their maintenance management system. All the reviewed papers are based on real case studies in different industries. However, the data gathered for those are from construction heavy machinery. To the best of our knowledge, this is the only research that is based on port and terminal load handling equipment. The main difference between logistic equipment is their features, the type of load, and the way they carry their cargo. These affect the performance and depreciation period of machines, which has a significant impact on the maintenance and expenses caused by that. As a conclusion, we try to propose a methodology that can forecast the optimum price for the maintenance cost of cargo handling machines used explicitly in ports and terminals, considering other research with most compatibility to this work.

3. METHODOLOGY

The process of digging through the data to extract patterns, knowledge discovery, and predict future trends can be implemented by three different scientific approaches: statistics, artificial intelligence, and ML (Data mining 2020).

Statistics based solutions study numerical relationships in data. Artificial intelligence solutions try to solve problems by applying human-like intelligence algorithms on data. Machine learning solutions are taking advantage of mathematical-based models to learn from data and make predictions.

In this chapter, some of the most popular data mining methods will be explained as possible solutions to the raised issue in this research.

3.1 Machine learning (ML)

Based on the use case, different definitions are applied to machine learning. Tom M. Mitchel (1997) defined machine learning as a study of the computer algorithms being capable of improving themselves automatically by learning from experience.

There are two types of techniques available in machine learning, Supervised and unsupervised algorithms. **Supervised** ML models are trained by some examples of input-output pairs (labeled data) to learn the mapping function between input variables of *X* and their output variable of *Y*: $Y = f(X)$. In supervised learning, the output value is available directly for the model (Russell and Norvig, 2009).

However, not all datasets are labeled by corresponding outputs. In these cases, **unsupervised** ML models will be offered to recognize probable patterns or structures in the data, while no explicit feedback is supplied (Russell and Norvig, 2009).

In the context of this research, supervised machine learning algorithms are considered as a method of data analysis to estimate and predict the future behavior of the maintenance cost of the fleet by learning from historical data.

3.2 Linear regression (LR)

Linear regression (LR) is a supervised machine learning technique. It approaches the problem by fitting the best linear function of inputs to their outputs. The task of linear regression is defined as bellow:

By having a training dataset which contains N pairs of input-output examples as $(x_1, y_1), (x_2, y_2), \dots (x_N, y_N)$, each numerical output of y_j generated by an unknown function $y = f(x)$. Linear regression learner is discovering the best function h among all the other hypothesis functions that approximate the true function f. In order to measure the accuracy of the hypothesis h, we test it with a set of samples that are distinct from the training set. If the function h was able to predict the average value of y for each input, we assume that hypothesis generalized well (Russell and Norvig, 2009).

If we have one explanatory variable of x in exchange for a dependent variable y, then the linear regression that explains the relationship between them is called **univariate linear regression**.

A univariate linear function follows a line equation of $y = w_1 x + w_0$. The coefficients of this equation are called weights and will be found during the learning process. Finding the weights in a way that minimizes the squared loss, L_2, gives the equation of the line that fits the data. Line equation is defined as bellow:

$h_w(x) = w_1 x + w_0$,

where weights will be defined as the vector $[w_1, w_0]$. The lag between the predicted and actual value of the dependent value is calculated as follow:

$$Loss(h_w) = \sum_{j=1}^{N} L_2\left(y_j, h_w(x_j)\right) = \sum_{j=1}^{N} \left(y_j - h_w(x_j)\right)^2 = \sum_{j=1}^{N} \left(y_j - (w_1 x_j + w_0)\right)^2$$

Minimizing the result of the above equation provides the best possible estimation for our model. This is possible by calculating the derivatives of the above equation with respect to w_1 and w_0 , and equating them to 0. The equations are as follows:

$$\frac{dh_w}{dw_0} = -2\sum_{j=1}^{N}(y_j - w_1 x_j - w_0) = 0$$

$$\frac{dh_w}{dw_1} = -2\sum_{j=1}^{N}(y_j - w_1 x_j - w_0)\, x_j = 0$$

Solving the system of above equations provides the values of weights as follows:

$$w_1 = \frac{Cov(x,y)}{Var(x)} = \frac{\sum_{j=1}^{N}(y_j - \overline{y})(x_j - \overline{x})}{\sum_{j=1}^{N}(x_j - \overline{x})^2}$$

$$w_0 = \overline{y} - w_1 \overline{x}$$

Multivariate linear regression is almost similar to the univariate case. Each sample of x_j in the multivariate case is a vector with n elements. Instead of a line, a hyperplane should be found to fit the output while minimizing the error calculated according to the loss function. The equations of the hyperplane space and the weight vectors that minimize the value of the loss function are as follow:

$$h_{sw}(x_j) = w_0 + w_1 x_j^{(1)} + w_2 x_j^{(2)} + \cdots + w_n x_j^{(n)} = w_0 + \sum_{i=1}^{n} w_i x_j^{(n)}$$

$$w^* = \underset{w}{\operatorname{argmin}} \sum_i Loss(h_{sw})$$

$$Loss(h_{sw}) = \sum_{j=1}^{N} L_2\left(y_j, h_{sw}(x_j)\right)$$

The minimum value of the loss function can be calculated by either analytical solutions or gradient descent (Russell and Norvig, 2009). In order to explore the strength of the relationship between each independent variable with CMC's (dependent variable), different univariate linear regression models are implemented in the next chapter.

3.3 Regression trees

Same as a linear regression model, decision trees are a method of supervised learning algorithms. A vector of independent attributes is a decision tree function input, and a single output value will be returned as the decision. Decision trees work with both discrete or continuous data points. The leaf nodes of the tree represent the outputs or, in other work, the decision (Russell and Norvig, 2009).

The conventional algorithm to construct a decision tree is top-down. A variable with more capability on splitting the set of items will be chosen on the upper layers of the tree (Rokach and Maimon, 2005). Different metrics are available to measure the homogeneity of the target attribute within the subset, such as Gini impurity. Regression trees are a type of decision tree and are used when the real number for the predicted value can be considered, such as prediction of the cost for the service work orders.

3.4 Gradient boosting regressor (GBR)

Gradient boosting is a supervised ML technique and one of the widely used ensemble methods. To understand GBR first, we need to know the idea behind boosting. In this method, we assume each example of the training set has an associated weight of more than zero. Examples with higher weights play a more critical role during the learning process of a hypothesis. All examples are given the same weight in the first step of boosting and from this dataset the first hypothesis, h_1, will be generated. Training data

records are classified correctly and incorrectly with this hypothesis. The goal of the next hypothesis is improvements in the classification of training examples set in comparison with the previous one (Russell and Norvig, 2009).

GBR makes predictions with a single strong learner, which is built as a combination of several fixed-size weak learners. A weak learner is a small tree that its performance is slightly better than a random guess. Decision trees have the ability to handle data of the mixed type and to model complex functions which make them valuable for gradient boosting models.

The goal of gradient boosting regressors is teaching the predictive model of $F(x)$ by minimizing the Loss function. Loss function or cost function is a function to evaluate the accuracy of our model. To improve the model's accuracy, the value of the chosen Loss function should be minimized by an optimization solution. A single additive model of gradient boosting regressors form as follow:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(\mathrm{x}), \qquad (10)$$

in which $h_m(\mathrm{x})$ is the latest week learner with the ability in the prediction of pseudo residuals, and It forms the following equation:

$$h_m = \underset{h}{\operatorname{argmin}} \sum_i L(y_i, F_{m-1}(x_i) + h(x_i)).$$

Attributes that form trees are chosen by the best split points calculated with purity scores like the Gini index. The first step toward making a GBR is to initialize a leaf with a constant value by the following equation and next is to expand it incrementally and greedily:

$$F_0(x) = \underset{\gamma}{\operatorname{argmin}} \sum_i L(y_i, \gamma), \qquad (11)$$

where $L(y_i, \gamma)$, is the Loss function, y_i is the observed value, and γ is the predicted value in the first step. The default choice of the loss function for regression problems is usually least squares, and the sum of the derivatives of it initializes a constant value for the precited values of the first step equal to the mean of target values.

To solve the minimization, gradient boosting uses the steepest descent, which is the negative gradient of the Loss function, as shown in equation 11. By taking derivatives of the Loss function with respect to the previously predicted value and set the sum of the derivatives equal to zero, the minimization problem will be solved numerically:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}, \qquad (12)$$

where r_{im} is the pseudo residual that calculated for the i^{th} example for the m^{th} tree, and $F(x)$ is predicted values of the latest week learner in equation 12. After calculating the

residuals for all examples available in the dataset, a weak learner will be built to model the pseudo residuals and parameterizing the tree with functional gradient descent to create its terminal regions, R_{ij}. For tree's regions in the weak learner a separate output value of γ_{jm} calculates as below (Friedman, 1999):

$$\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma),$$

where j is the leaf index, m is the tree's number and i is the index of the samples. To improve the overall output of the model, the output of each tree will be added to the existing sequence of trees outputs.

$$F_m(x) = F_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} \; I_{R_{jm}}(\mathrm{x}),$$

The summation in the above equation means all regions in which sample x can be found should be added up. Regularization by shrinkage is an important part of the implementation of the gradient boosting methods to reduce the overfitting effect. In the above equation, parameter ν, is a number between 0 and 1 and called the learning rate. Using a small learning rate ($\nu < 0.1$) yields dramatic improvement in the model with the cost of increasing its iteration, complexity, and computational time. The contribution of each tree to the model is weighted by the learning rate (Friedman, 1999).

GBR solves the problem by taking small learning steps from the average of output as the initial prediction toward the true values. Although they are hight in computational time, they are a great fit for the purpose of this research since they can manage mixed data types contains numerical and categorical types.

3.5 Model Validation methods

The closeness of the model mimicking the actual system definition is one of the biggest concerns during the process of building a statistical model. The process that confirms the reliability of the model's result concerning the real system output, is referred to as model validation. What we are trying to reach is a high degree of validity and an accurate representation of the system. From this viewpoint, validating a model is the task of verifying that the model is adequately accurate for the predetermined application, in the domain of its applicability (Schlesinger, 1979). A model that is built on a specific application might not be valid for some other applications therefore it can be validated for a specific domain only.

There are three practical difficulties to reach the validity of a model: 1) Data shortage, 2) Lack of control over the real system's input variables, 3) the ambivalence of the underlying correlation and probability distributions for both real system and created model (Kotz et al., 2006). The simplest way of model performance estimation is to the resubstitution estimate of risk:

$$R_{resub}(\emptyset) = \frac{1}{n} \sum(y_i - \emptyset(x_i; \tau))^2,$$

where $R_{resub}(\emptyset)$ is the average of the squared residuals of the training samples. However, the underestimation of the risk by a substantial margin is a drawback of this method. This behavior is called overfitting since the model was built in a way to fit the dataset too well (Stuetzle, 2005).

A popular approach to calculate the risk is by extracting a test set from the initial dataset. In this way, the model is trained by part of the dataset, called the train set, and the average loss function is calculated when the model makes the prediction for the test set. However, depends on the randomness of the samples in these datasets, there is still the possibility of overfitting because of the knowledge leaking from the test set into the model. Besides, we may not access such a big dataset to be able to split it into training and testing datasets (Stuetzle, 2005).

3.5.1 Cross-validation (CV)

Cross-validation is considered as the solution for the mentioned problems and it is one of the key methods of model performance assessment (Stone, 1974). Cross-validation categorizes under the sample reuse models since the training dataset is randomly divided into k number of equal size subsets $\tau_1, \dots, \tau_k$. The computational efficiency of this approach differs based on the number of subsets. Bigger K is causing a computationally expensive process. However, it cannot be a matter based on available resources in this research. Based on empirical evidence, suitable k is considered a number between 5 to 10 subsets (Stuetzle, 2005). Figure 7 is showing a general data division in 5-fold cross-validation.

Figure 7. *Illustration of 5-fold cross-validation*

This structure is called **k-fold CV** and can be described best in an algorithmic manner. Following procedure is repeating for each "fold" in τ^{-i} as training dataset while i-th subset is removed:

For $i = 1 \dots k$ {

- Generate a prediction model $\emptyset(x, \tau^{-i})$ based on the training samples.
- Compute the Loss L^i by testing the model with i-th subset:

$$L^i = \sum_{j \in \tau^i} (y_j - \emptyset(x_j, \tau^{-i}))^2$$

}

The risk estimated by K-fold CV is the average of the Loss values calculated in the loop as below: $\bar{R}_{cv}(\emptyset) = \frac{1}{n}\sum_{i=1}^{k} L^i$.

3.6 Hyperparameter optimization methods

In order to achieve the best cross-validation estimator, it is recommended to optimize the hyperparameters of the estimator. Hyper-parameters are parameters that cannot be learned directly within estimator therefore their values have to be set before starting the

learning process (scikit-learn, 2020). Tuning hyperparameters returns a tuple of hyperparameters with the ability to provide an optimal model which minimizes a predefined loss function on given data. Searching for optimal hyperparameters is commonly performed manually. A large number of estimator's hyperparameters make this process impractical. Due to such flaws, the idea of automatic approaches in hyperparameter optimization received a big amount of attention. However, depends on the available computational resources, the nature of the desired estimator, size on ensembles, and the size of the dataset each evaluation may take considerable time. Therefore, there is a need for finding an efficient tuning process of hyperparameters that require minimum amount of objective function evaluation (Claesen and De Moor, 2015).

The two most general approaches to solve the hyperparameters search problem are **exhaustive grid search** and **random search**. Grid search is building a model for each combination of listed values of hyperparameters exhaustively to evaluate each model and pick a learning algorithm with high accuracy. A performance metric solution such as cross-validation or held-out validation on the training set is needed to guide the grid search algorithm (Hsu, Chang, and Lin, 2003). Picture 8 is showing the visual representation of the grid search.

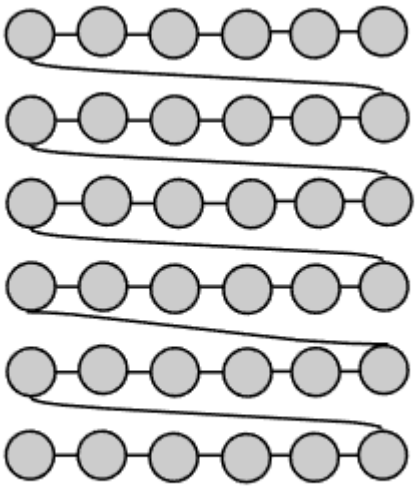

Figure 8. *Grid search visual representation*

On the other hand, the random search selects hyperparameters randomly instead of exhaustive enumeration (Ghawi and Pfeffer, 2019). While grid search is suffering from the curse of dimensionality, random search takes less computational time than grid search and may perform better when a small number of hyperparameters involve in the process (Bergstra and Bengio, 2012).

Despite decades of research into hyperparameter optimization approaches, there are several reasons that grid search is preferred in various research:

- The simplicity of its implementation
- As it creates a grid of hyperparameters values and enumerates all possible combinations of hyperparameters it typically finds more reliable values combination.

3.7 Coefficient of determination (R Squared)

Coefficient of determination is a statistic that assesses how well is a linear regression model in explaining and prediction the dependent variable based on the input data. More specifically, it is the percentage of the dependent variable (y) variation that is explained correctly by independent variables (x). This measure can be calculated with the following formula:

$$r^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum(y_i - f_i)^2}{\sum(y_i - \bar{y})^2},$$

where *RSS*, the residual sum of squares, is the total squared of errors between the dependent variable (y_i) and the predicted values (f_i). *TSS* is the total variation in y and it is defined as the sum of all squared differences between the observations (y_i) and their overall mean ($\bar{y}$). Generally, the closer value of r^2 is to 1, the better fit of the model to the data. Any r^2 value bigger than zero means that the regression analysis predicts the target variable better than just using a horizontal line through the mean value (Nerdy, 2020).

3.8 Practical implementation

The aspects of the study that concerns statistical analysis and machine learning methods were carried out using python programming and modules of scikit-learn library. The Jupyter notebook provided in Amazon SageMaker was the open-source web application to document and share the live code related to this study. Tableau and Minitab are the analytical platforms used for some of the data visualization processes. SQL queries are created to collect the relevant data from Athena AWS. And the Amazon S3 buckets were used as the main storage to read and manipulate the queried data later on.

4. EXPERIMENTS

As it is indicated in previous chapters, a suitable repair cost forecasting model leads to achieving a cost-effective system. Finding the influential and optimal parameters on maintenance is all the company's goal since it could help to manage the cost system while keeping the customers satisfied. In this chapter, a methodology is proposed to forecast the optimal repair cost based on the machine's usage. Moreover, we are tracking the most influential parameters.

The overall AI/ML architecture, followed by the partner company, can be summarized in two main phases. First is data exploration, which data staging and preparation process are applied. Secondly, model lifecycle management, which breaks to the implementation of feature engineering, model creation, training and inference, and model exposure. Figure 8 is showing the sequence of this process.

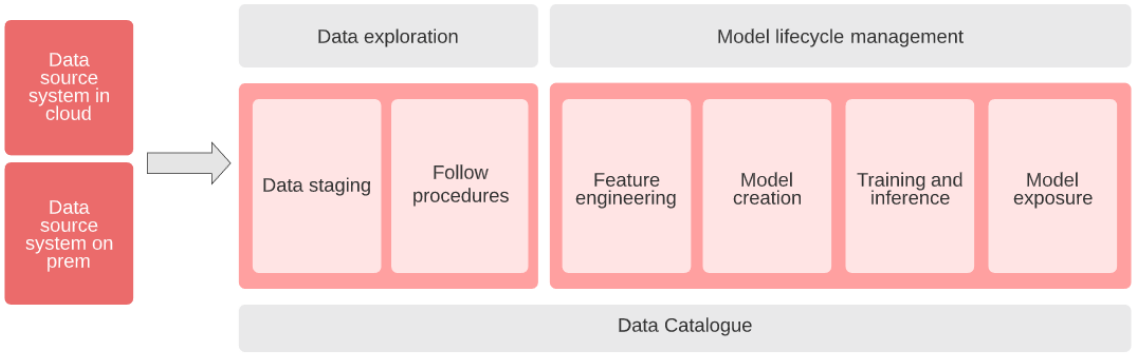

Figure 9. *Cargotec Conceptual AI/ML Architecture*

The implementation of this research took advantage of the resulted dataset of the first stage, data exploration. In feature engineering, the process of quality assessment, preprocessing, and feature aggregation was implemented in which data was cleaned from the apparent anomalies or incomplete records, and meaningful features were created. In model creation, where basic trend analysis was performed to visualize the actual trend of the maintenance cost, cumulative cost modeling was conducted. Lastly, the data mining model was created. The next step was training the model with three different data approaches. The results are exposed in the last phase, and model evaluation and validation took place in this last step by comparing the result, and the best algorithm was selected. Further, these parts are discussed in the same order.

4.1 Preliminary analysis

Before going through any pre-processing, browsing the distribution of RS's maintenance cost over time can give a better perspective on the cost behavior. Figure 10 is showing the frequency of the maintenance cost of all RS's in five sequential periods. Each period contains eight months of the machine lifetime.

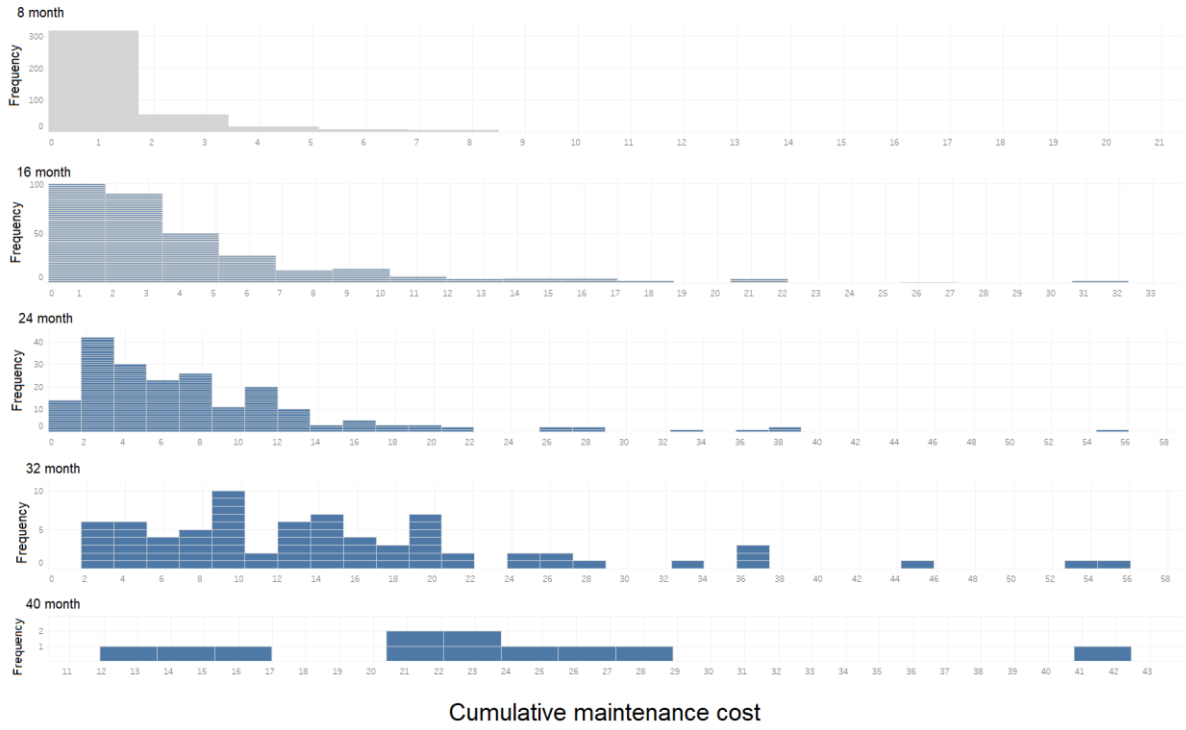

Figure 10. *SWO cumulative cost distribution for all available RS's in every 8 months of their lifetime.*

As can be seen from the picture, the CMC of almost all RS's is near zero in the first period. The distribution stays steady and right-skewed in the second period and the third period. However, the cumulative cost distribution fluctuated after two years of RS's life, and the right-skewed pattern of the diagram disappeared over time.

The diagrams also show some cases of CMC at the tails of the slopes, which may seem like outliers. However, their considerable SWO costs may be caused by the use-case environment and workload of the intended RS, which vary from one equipment to another. These make the prediction of the maintenance cost inaccurate by the traditional analyzing methods. Therefore, we would like to check if ML models can provide accurate answers.

In order to implement any ML model, data should be transformed, encoded, or processed in a way that machine be able to parse it properly. The following steps provide these characteristics for our case study dataset.

4.1.1 Data quality assessment

Inconsistency and missing values are common among field datasets due to flaws in the data collection process, data entry errors, or limitations of measuring devices. Such variations must be found and solved. In this case study, some similar outliers were encountered, as illustrated in Figure 9. The records of three different RS's were picked as an example of missing values. Each color represents an RS. As can be seen from the picture, these three RS's have increasing engine hours. It means they were under a workload, and when a piece of equipment is working, not only the number of engine hours is increasing, but also the amount of some other variables expected to increase. However, in this specific example, other cumulative attributes like the overall amount of fuel used, number of lifts, and kilometers traveled by the equipment resuming from 0 at some service work order date.

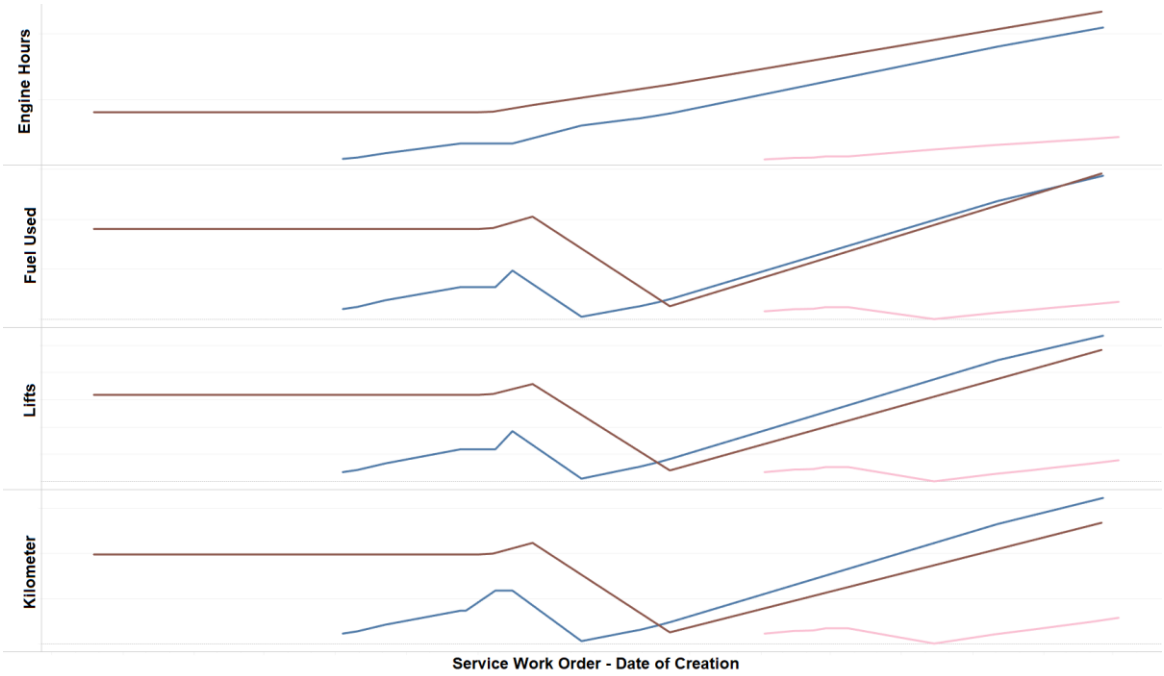

Figure 9. *Abnormal drops on cumulative attributes of three RS's.*

This anomaly was resolved by adding the new values of features with the last valid value of that feature before resetting.

Some features play a crucial role in prediction and the accuracy of the model. The value stored in these entities should be reliable, and missing value may cause side effects on the results. The elimination of data records with missing values in critical entities was a safe and effective strategy to cope with this problem.

4.1.2 Feature exploring

In this section, we try to put the data in a better perspective and offer a high-level view of the data in order to come up with an ideal final dataset to feed our machine learning model. To achieve that, first, we observe the behavior of CMC in relation to changes in some critical attributes. As mentioned in chapter 2, we are expecting a growing CMC by increasing the engine working time (See section 2.2). Figure 12 sampled the service work order of eight different RS's with almost the same distribution in their engine hours which follows Vorster (1980) pattern. The fitted trend line with **0.50** r^2 was the most descriptive of the data in comparison with the exponential trend line with a **0.48** r^2 and logarithmic trend line with **0.42** r^2.

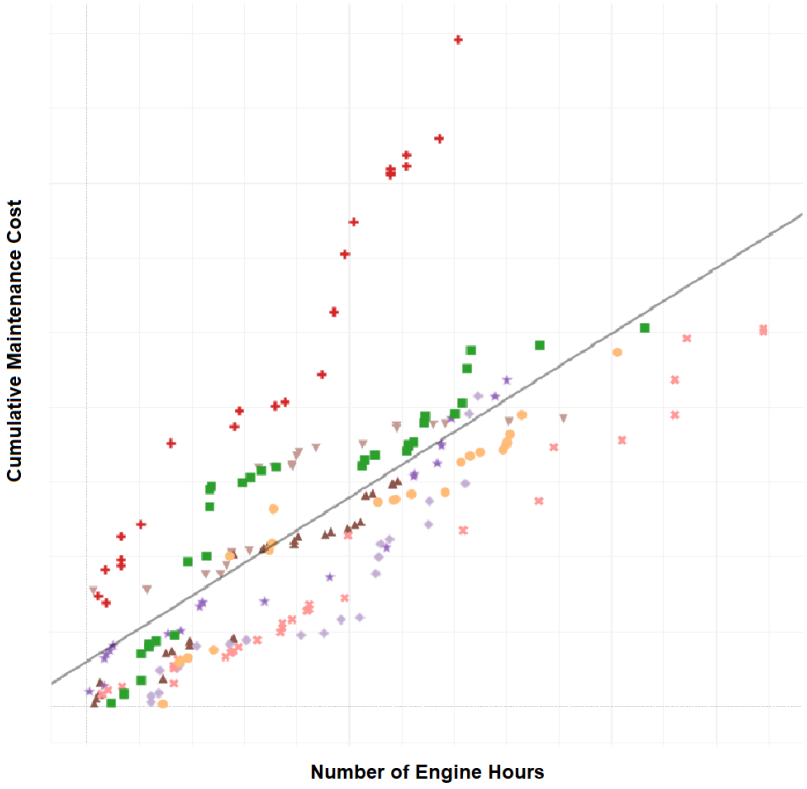

Figure 10. *CMC per number of engine hours for eight RS's.*

However, the behavior of CMC did not perform the same for all the sample equipment as we were expecting. Figure 13 is the cost of the service work order of all RS's over their engine working time to check if it follows the mentioned pattern (See section 2.2). Each RS has a unique color and shape. The exponential line stated before by Vorster (1980) gives a **0.23** r^2 in this dataset, which provides a weak estimation of the CMC for our RS's.

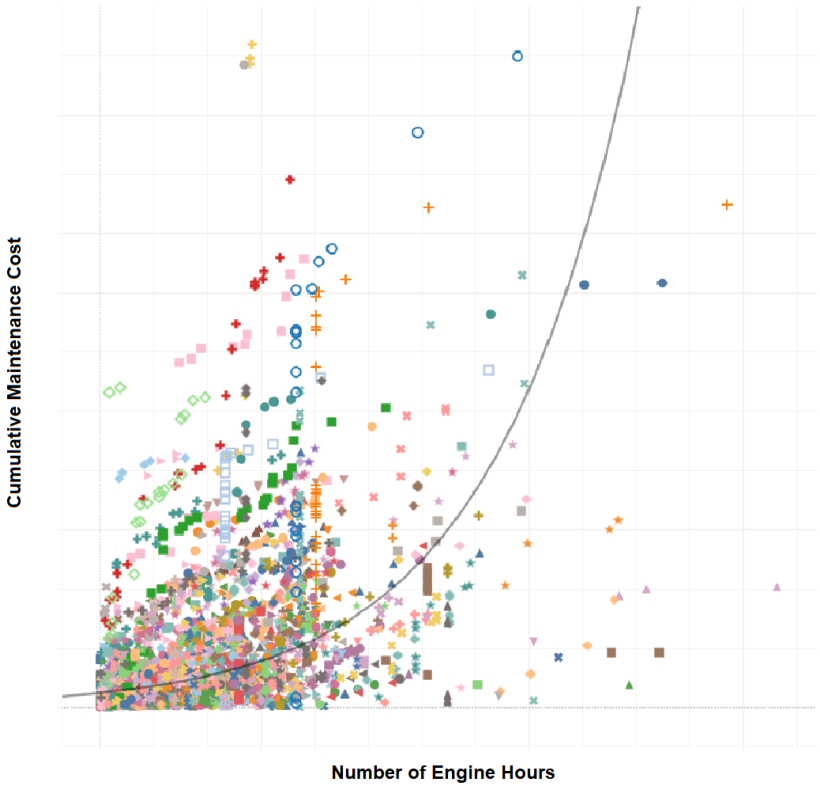

Figure 11. *Implementation of Vorster model for all RS's based on engine working hours*

Figure 14 highlighted some cases that their maintenance costs increased while the number of working hours did not change for some sequential work orders. It can be caused by several reasons such as misdiagnosing of the machine problem by maintenance experts or the need for replacing various spare parts to make the machine available again. In addition, there are machines with massive workloads and small maintenance costs; figure 15 highlighted some of the cases. These two figures are showing some examples of the equipment that does not follow the Vorster suggested pattern (CCM).

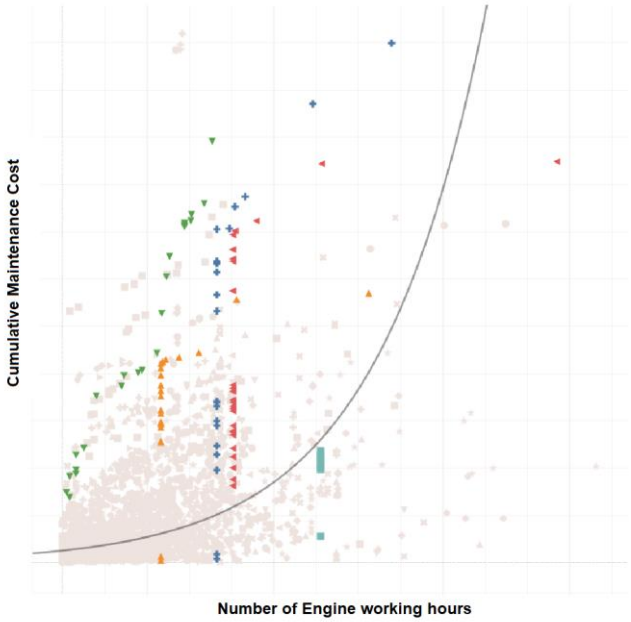

Figure 12. *Grows in maintenance cost of RS's with shutdown engines*

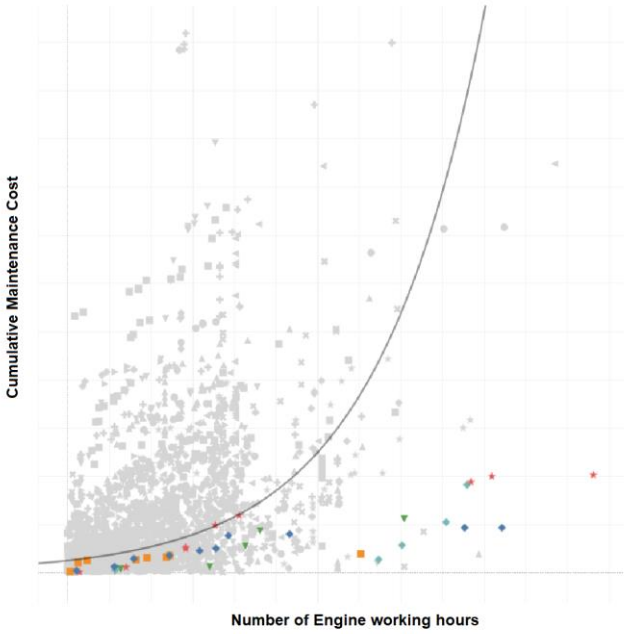

Figure 13. *RS's that have been working for a long time without needs for expensive maintenance*

The working environment, operator skills, equipment age may seem the only influential factors on the variation of the EMC. However, we found that unlike typical machines, the number of lifts is the other influential factors on EMC, and engine working hour is not the only measure of the workload of load handling equipment.

Figure 16 emphasis the fact that the number of cargoes that machine lifts and the number of its engine working hours are positively correlated with **0.89** strength. Each circle represented a machine, and the size of the circle represents the amount of its overall maintenance cost.

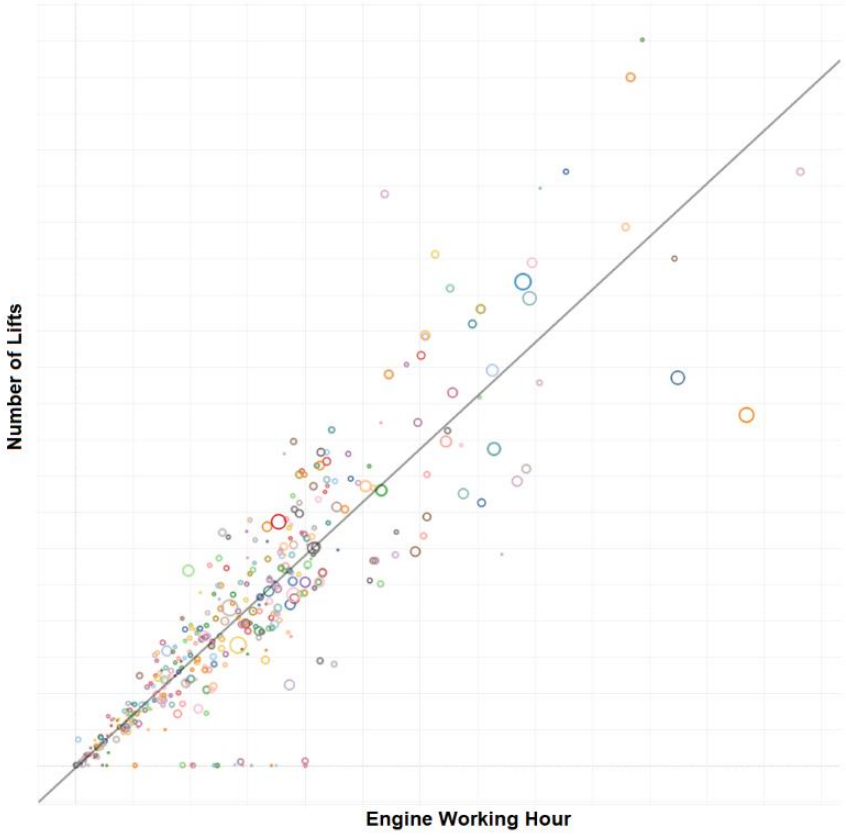

Figure 14. *Correlation between the number of lifts and engine working hours*

Figure 17 is showing the behavior of the CMC in comparison with the number of cargo that machines lifted. As can be seen from the scatter plot, CMC follows almost the same pattern as it did with engine hours. If we consider the number of the lifts as the measure of the equipment's age then the best fitted exponential line which is the implementation of Vorster method (CCM) gives a poor estimation of CMC with **0.19** r^2 for the available data.

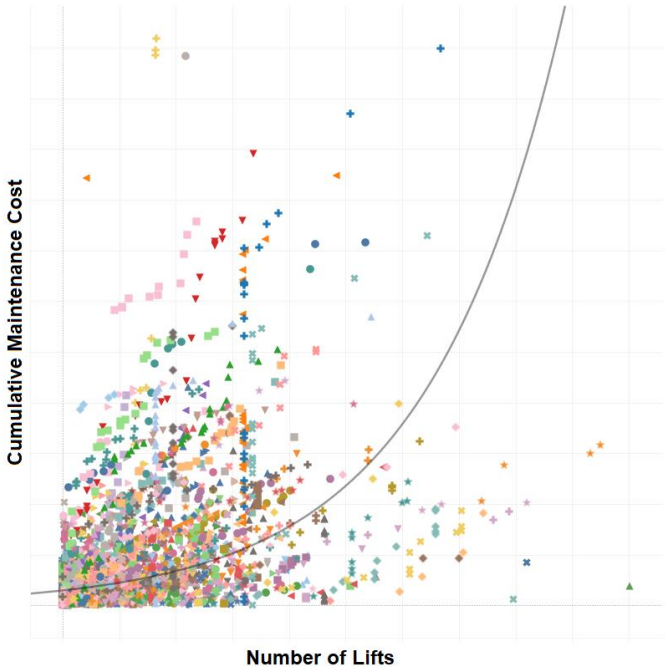

Figure 15. *CMC per number of lifts for all RS's*

Based on field expert's statements (Josse, 2017) and what we browse through the data, the following variables are the most influential attributes on the cost of a maintenance work order, and they are detailed in Table 1.

Table 1. *Influential variables on CMC*

Variable name	**Description**
Engine working hours	The number of hours that machine engine worked
Lifts	Number of freight containers handled also called "Number of lifts"
Kilometer	Number of Kilometre distance the machine traveled
Tons	Total weight handled in tons
Fuel	The amount of fuel in litters that machine consumed
Engine	The type of machine's engine
Transmission	The type of machine's transmission

Based on researches mentioned in Chapter 2, commonly used performance indicators derived from given attributes calculated based on engine working hours. It is detailed in Table 2.

Table 2. *Variables of interest retrieved from Table 1 based on engine working hours.*

Variable name	Calculation	Description
Meter per engine hour	Kilometer*1000/Engine hour	The distance that equipment traveled per engine hour.
Ton per engine hour	Ton/hour	The weight handled in each hour of the engine working.
Fuel per engine hour	Litter/hour	The amount of fuel machine uses per engine working hour.

It is straightforward to assume that other numerical attributes (fuel used, distance, and cargos weight) will be positively correlated with the engine working hour, the same as the number of lift.

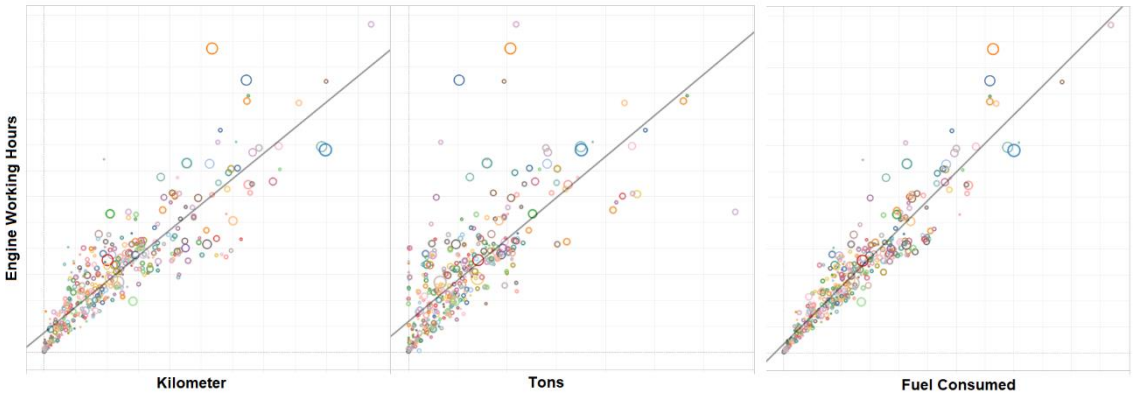

Figure 16. *Correlation of Engine working hours with cumulative attributes (Kilometre, Tons, fuel consumption)*

Figure 18 is showing the behavior of these attributes toward engine hours. It can be seen from the picture that a linear relationship between cumulative attributes and engine working hours is available. For example, it is technically assumed that two hours of engine working consume a double amount of fuel than one hour of engine working (Josse,

2017), and it can explain the strong positive correlation between fuel consumption and engine working hours.

However, we already found that the number of lifts that a machine has done is another measurement for its workload (see Figure 17). Therefore, the following attributes can be derived, as shown in Table 3.

Table 3. *Variables of interest calculated based on the number of lifts*

Variable name	Calculation	Description
Meter per lift	Kilometer*1000/lift	The distance that equipment traveled per lift.
Ton per lift	Ton/lift	The weight is handled by machine in each lift.
Fuel per lift	Litter/lift	The amount of fuel machine uses per lift.

After going through preprocessing phases and filtering out unusable data the resulted dataset shrunk from 5000 observations to 3000. The remained samples are coming from machines with accurate details in all relevant attributes. To discover rather the engine hours factor or the number of lifts factor is more correlated to the amount of maintenance cost two final datasets were created based on the raw data. The attributes of each dataset are shown in Tables 4 and 5. Attributes that form Table 4 are called cumulative maintenance cost per lifts (CMCPL) dataset and attributes that form Table 5 are called cumulative maintenance cost per engine hour (CMCPHR) dataset.

Table 4. *Attributes in final dataset derived based on the number of lifts*

Variable name	Description
Meter per lift	The distance that equipment traveled per lift.
Ton per lift	The weight is handled by machine in each lift.
Fuel used per lift	The amount of fuel machine uses per lift.
Engine hours	The hours of the machine’s engine working.
Cost per lift	Maintenance cost per lift. (dependent variable)

RS Model	The model of the RS machine which differs based on the machine's usage, strength, and ecosystem.
SWO day duration	The length of the service work order in days

Table 5. *Attributes in final dataset derived based on engine working hours*

Variable name	**Description**
Meter per engine hours	The distance that equipment traveled per engine hours.
Ton per engine hours	The weight is handled by machine in each engine hours.
Fuel per engine hours	The amount of fuel machine uses per engine hours.
Lifts	The number of lifts the machine has carried.
Cost per engine hours	Maintenance cost per engine working hours. (dependent variable)
RS Model	The model of the RS machine which differs based on the machine's usage, strength, and ecosystem.
SWO day duration	The length of the service work order in days

Figure 19 is showing the gaussian estimate of the probability density function (PDF) of the maintenance cost for the two final data sets. The red dashed line represents the median values of cost per lift distribution and the blue line showing the median for cost per engine hour distribution. The diagram shows that medians for both distributions are very close to each other. However, the cost per engine hour distribution has a larger variance.

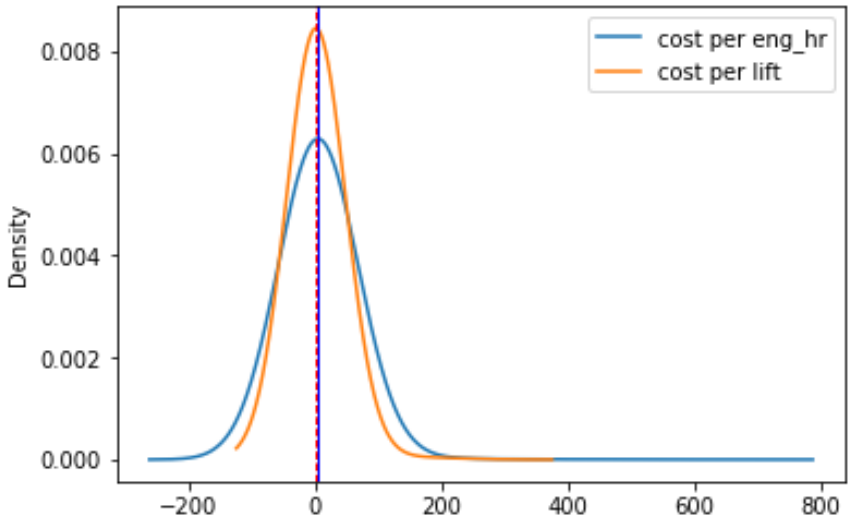

Figure 19. *The probability distribution function of 2 datasets*

4.2 Data Division

Dividing the data is the first task before model creation. The datasets are dividing into training, validation, and test sets. The training dataset is used to train the model, the validation sets are used for evaluation and verification of the model during iterative training to choose the best meta parameters. Once the hyperparameters have been selected and the final model was built and trained, we can evaluate the performance of this ultimate model with the test set.

Since the size of the dataset was too small, only 0.1 part of it separated randomly as the test set. By random sampling of the rest of the data into three equal-size partitions, one partition is assigned to the validation set and two others as a training set in each round of the cross-validation process.

4.3 Linear regression

To investigate the distribution properties, an LR model (see section 3.2) was applied to all the observations in the data. Furthermore, we were interested to see how well the datasets could be fitted by a linear model. Plots in Figures 17 are describing the residuals (the difference between predicted values with true values of CMCPHR) versus predicted CMCPHR goodness of the linear regression fit on the CMCPHR dataset as well as actual values of CMCPHR versus their predicted values. As shown in the left plot in Figure 17 linear model can not successfully express the underlying relations in the data with -0.09 r^2. In the plot on the right, each point is a prediction of CMCPHR made by the LR model on the x-axis and the accuracy of the prediction is on the y-axis. The distance from the

line at 0 is showing how bad the prediction is for the actual CMCPHR value since $Residual = Observed - Predicted$.

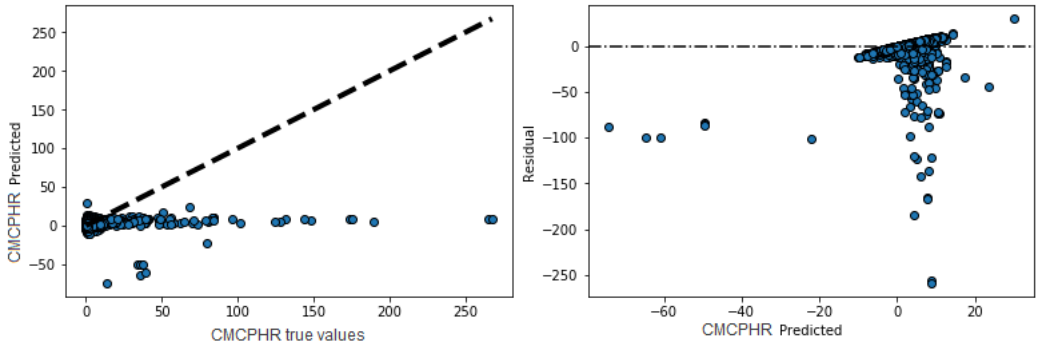

Figure 17. *Predicted values versus the observed values for CMCPHR by the LR model on the left and residuals vs predicted values of CMCPHR on the right*

Figure 18 is showing the same plots for CMCPL datasets. As can be seen from the left plot LR model was more accurate in predicting CMCPL with -0.1 r^2, however, it is still a very poor estimation since the value of r^2 is negative and it simply means that using the mean value of the actual values for predicting the future CMCPL will give better accuracy than the LR model.

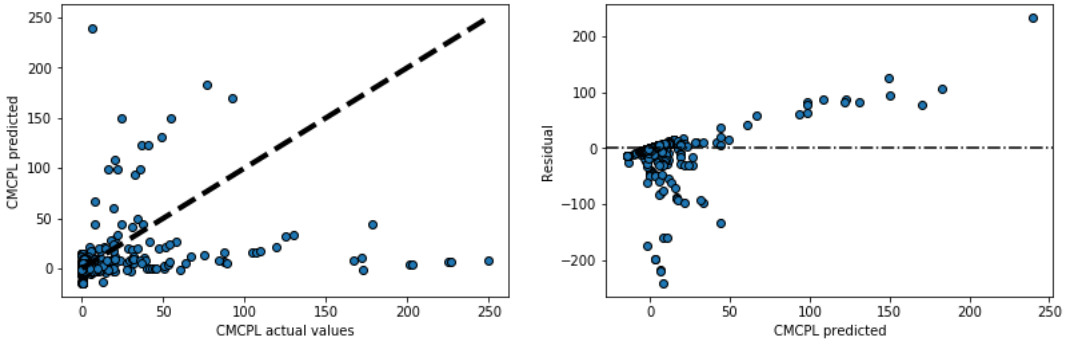

Figure 18. *Predicted values versus the observed values for CMCPL by the LR model on the left and residuals vs predicted values of CMCPL on the right*

4.4 Gradient boosting regression

For fitting a gradient boosting regressor to the training dataset we used the Sklearn ensemble module in Python (sklearn.ensemble.GradientBoostingRegressor documentation, 2020). We used the data as described in the data division section (see 4.2). To

choose the best combination of hyperparameters for the final model, the Sklearn GridSerachCV module is used to implement an exhaustive search over specified parameter values. The use of grid search may seem too straightforward and naïve, however, to experience all the possible combinations of data this method is preferred over other types of hyperparameters tuning methods. Furthermore, the required computational time for this approach was not long based on the provided computing resources provided for this research. The sets of hyperparameters and the range of probable values for each one is shown in the following table.

Table 6. *List of tunned hyperparameters and the range of values*

Parameters to optimize	Range of values
Learning_rate	[0.01,0.015,0.025,0.05,0.1]
max_depth	[4,6,8,9,10]
max_features	[0.1, 0.3, 0.5, 1.0]
min_samples_leaf	[1,2,5,10]
n_estimators	[100,120,300,500,800]

The n-estimator parameter is the number of weak learners to fit in the model and it is strongly interactive with the learning rate parameter. As mentioned in the previous chapter, the lower amount for learning rate reduces the effect of the problem of overfitting, however, the increase of the number of weak learners might be needed to maintain a constant training error. In this way, we can obtain better test error (Hastie, Tibshirani, and Friedman, 2009). Max depth parameter is the depth of each weak learner and a split point at any depth will be considered if at least the minimum number of samples is reached in each branch. This minimum is applied by the “min_samples _leaf” parameter and may have a smoothing impact on the regression model. The fraction of the features at each split is indicated by the “max_feature” parameter.

After training the model with each dataset we achieved the best estimator with the smallest loss which is chosen by the grid search. The best hyperparameter values for the models trained for each dataset are detailed in the following table.

Table 7. *Best values for hyperparameters in each trained model*

Dataset	Learning rate	Max feature	n-estimator	Max depth	Min samples leaf
Data calculated based on Lift (CMCPL)	0.015	0.3	120	8	1
Data calculated based on engine hours (CMCPHR)	0.05	0.1	120	10	5

Coefficient of determination of regression function, r^2, was the metric used to calculate the accuracy of each model for training and test set. The best value of this metric is 1.0 while the worst case can be negative and the score of the model that always predicts the expected output value disregards the input features is equal to 0. As shown in Table 8 the accuracy of the model trained by the dataset calculated per number of lifts is more accurate on both training and testing sets and it shows the maintenance cost of each machine depends more on how many lifts they have done rather than the number of hours of their engine working.

Table 8. *Summary of* $\boldsymbol{r^2}$ *scores resulted from the GBR model*

Dataset	Training set $\boldsymbol{R^2}$ score	Test set $\boldsymbol{R^2}$ score
Data calculated based on Lift (CMCPL)	0.94	0.83
Data calculated based on engine hours (CMCPHR)	0.77	0.30

To visualize the accuracy of the models on each test set scatter plots of the predicted values vs actual values are shown for estimators of each test sets in Figures 19 and 20.

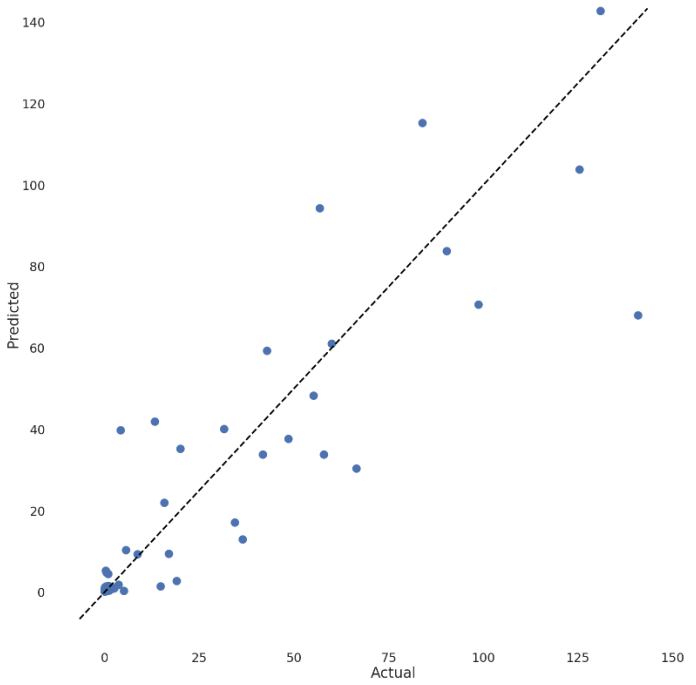

Figure 19. *Predicted values vs actual values in CMCPL test set*

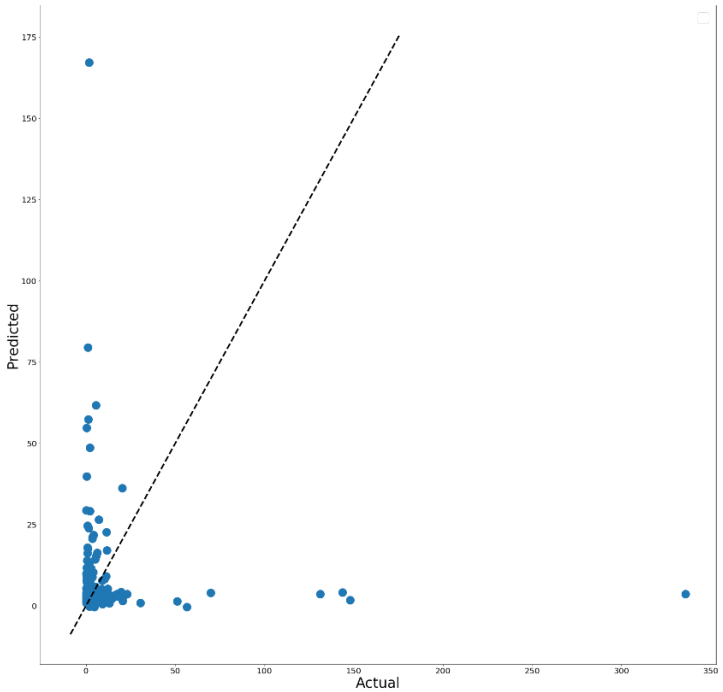

Figure 19. *Predicted values vs actual values in CMCPHR test set*

From the above scatter plots it is clear that prediction of maintenance cost based on the number of lifts is more accurate as the points are closer to the regressed diagonal line in comparison with Figure 19.

The final gradient boosting regressor was used to evaluate the relative importance of different features. The results of this rating are described in Figure 21.

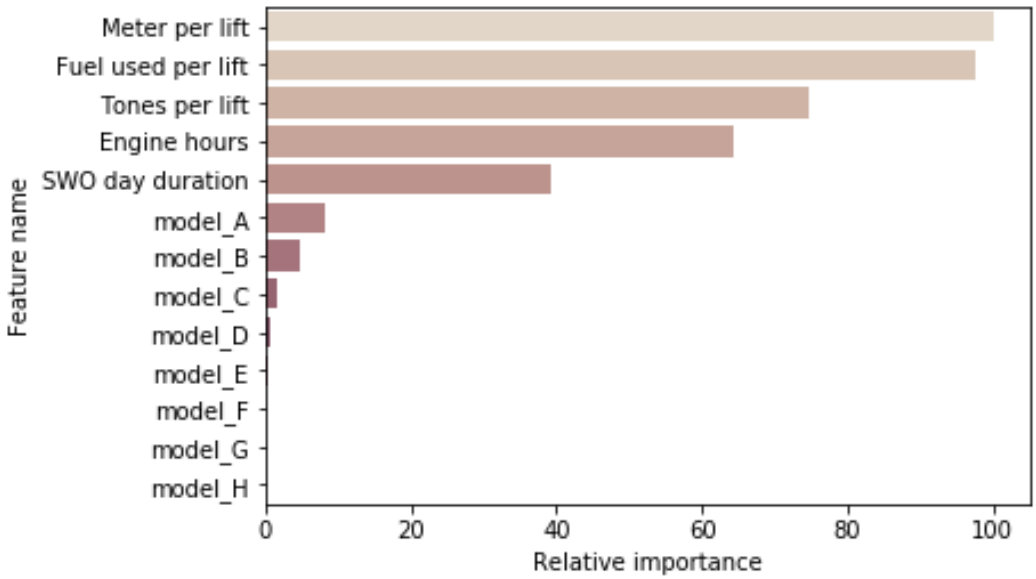

Figure 20. *The relative importance of features*

This chart is showing that the most important features for predicting maintenance costs are: meter per lift, fuel per lift, tons per lift, engine hours, and duration of SWO in days. This chart is showing that RS's models which emphasize the capacity of the machines do not come close in importance to features from the telemetry data.

5. CONCLUSION

5.1 Recommendations

To show changes in the learning performance of the GBR model for the CMCPL data set over time in terms of experience, following the learning curve on the train and validation datasets is plotted in Figure 22.

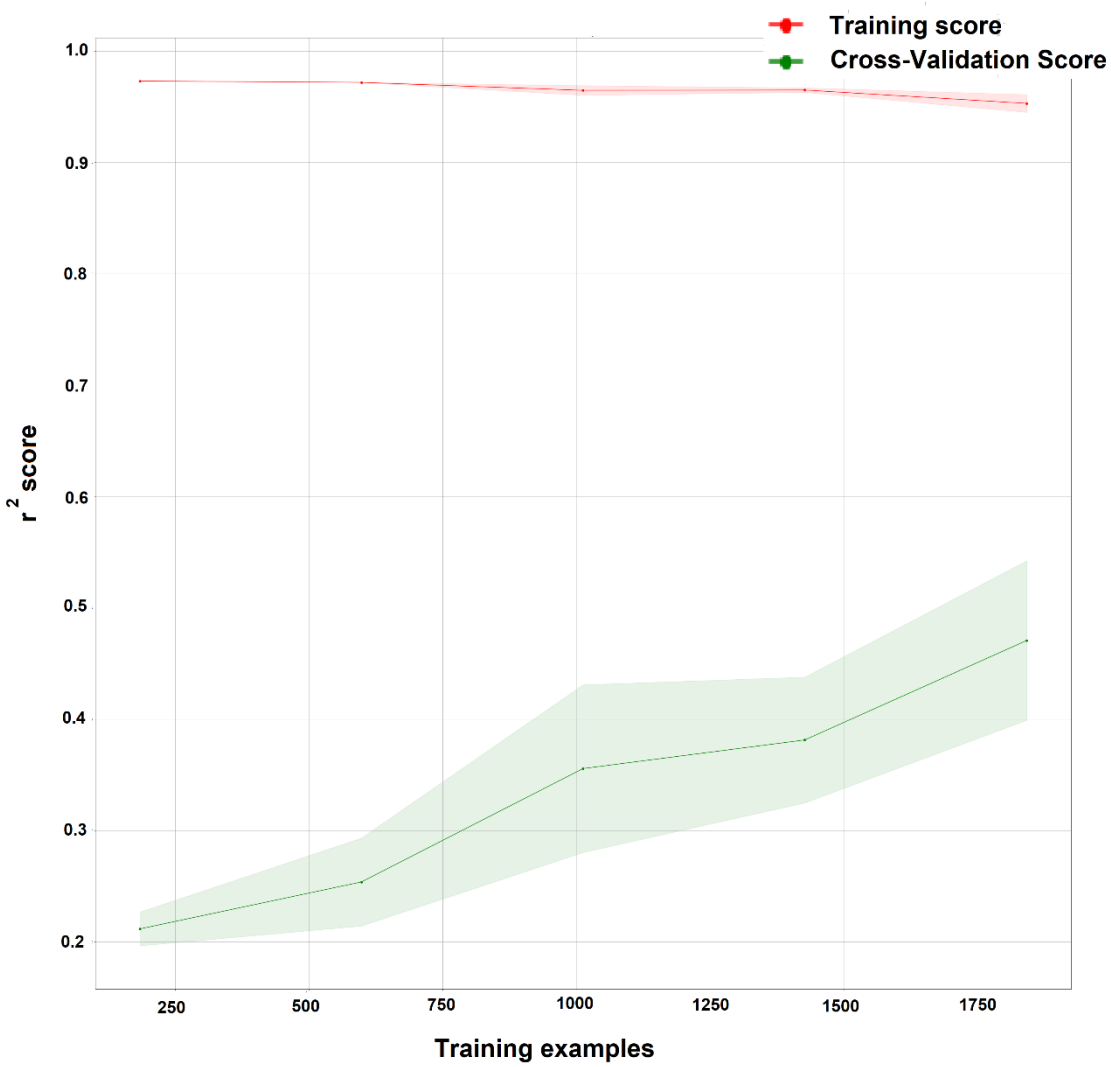

Figure 21. *Learning curve on training set calculated based on lifts*

The big gap between the cross-validation sets line and training scoreline shows that our model is suffering from overfitting due to the high variance in samples. An action we have identified that may solve this problem and improve the quality of the maintenance cost prediction is the increasing of samples. However, the possibility of improving the

model with the small size of the provided dataset is not achievable as this data is just a short glance of the whole lifetime of the equipment.

Another possible development would be to add attributes engine type and transmission to the data. The quality of data in these attributes was poor as there were a lot of machines available with no information about their engine and transmission types. Since we could not afford to lose more observations, the mentioned attributes omitted from the datasets as the result.

5.2 Final conclusions

Nowadays the engine hours of the machine are considered as the main indicator of machine maintenance cost in manufacturers. However, a comparison between the reviewed solutions in the second chapter with the result of this study indicates that the calculation of raw data based on the number of lifts that machine moves provide a better understanding of the cost of RS's service work order. This result makes sense since the lifts as an attribute have done a better measurement of the work than the hours taken to do the same work. Furthermore, the linear regression model is not a good representative of the data behavior while the gradient boosting regressor achieved good accuracy of r^2 score. The cross-validation method is used to partition the data for train and validation sets, however, the high variance in the small size dataset caused a bit of overfitting in the model. The most influential features for prediction are extracted from the data and the effects of duration of the service work order in the prediction appeared relevantly important and noticeable.

REFERENCES

Kalmar - Cargotec, 2019. Kalmar Hiab MacGregor Our businesses: Kalmar Hiab MacGregor. URL https://www.cargotec.com/en/kalmar/ (accessed 8.10.19).

Reachstacker, 2019. Wikipedia, the free encyclopedia.

Josse, V., 2017. Fuel consumption analysis on container handling equipment. KU LEUVEN.

Rouse, Margaret. “What Is Telemetry? - Definition from WhatIs.Com.” WhatIs.Com, 1 Sept. 2005, https://whatis.techtarget.com/definition/telemetry.

Cousineau, Marc. “A Guide to Mastering Maintenance Work Orders.” Fiix, 22 Oct. 2019, https://www.fiixsoftware.com/blog/work-order/.

Rouse, Margaret. “What Is ERP (Enterprise Resource Planning) and Why Is It Important?” Techtarget, 1 Aug. 2019, https://searcherp.techtarget.com/definition/ERP-enterprise-resource-planning.

Rouse, Margaret. “What Is SAP? Definition from WhatIs.Com.” Techtarget, 1 Aug. 2019, https://searchsap.techtarget.com/definition/SAP.

Robert L. Peurifoy, P.E.; Clifford J. Schexnayder, P.E., Ph.D.; Robert L. Schmitt, P.E., Ph.D.; Aviad Shapira, D.Sc. Construction Planning, Equipment, and Methods, Ninth Edition. ELEMENTS OF OPERATING COST, Chapter (McGraw-Hill Education: New York, Chicago, San Francisco, Athens, London, Madrid, Mexico City, Milan, New Delhi, Singapore, Sydney, Toronto, 2018). https://www-accessengineeringlibrary-com.libproxy.tuni.fi/content/book/9781260108804/toc-chapter/chapter2/section/section18

Yip, Hon-lun, et al. (2014) Predicting the Maintenance Cost of Construction Equipment: Comparison between General Regression Neural Network and Box–Jenkins Time Series Models. https://www.sciencedirect.com/science/article/abs/pii/S0926580513001921.

Vorster, M. (1980). “A systems approach to the management of civil engineering construction equipment.” Ph.D. thesis, Univ. of Stellenbosch, Stellenbosch, South Africa.

Mitchell, Z. (1998). A Statistical Analysis Of Construction Equipment Repair Costs Using Field Data & The Cumulative Cost Model. Ph.D. Virginia Polytechnic Institute and State University.

Mitchell, Z., Hildreth, J., and Vorster, M. (2011). Using the Cumulative Cost Model to Forecast Equipment Repair Costs: Two Different Methodologies. Construction

Engineering and Management, [online] 137(10), pp.2-4. Available at: https://ascelibrary.org/doi/10.1061/%28ASCE%29CO.1943-7862.0000302.

Strang, G., and Herman, E. (2016). Calculus Volume 1. Houston, Texas: OpenStax, https://openstax.org/books/calculus-volume-1/pages/4-4-the-mean-value-theorem.

Yip, H., Fan, H., and Chiang, Y. (2014). Predicting the maintenance cost of construction equipment: Comparison between general regression neural network and Box–Jenkins time series models. Automation in Construction, 38, pp.30-38.

Specht, D. (1991). A general regression neural network. IEEE Transactions on Neural Networks, 2(6), pp.568-576.

Friedman, J., 1999. Greedy Function Approximation: A Gradient Boosting Machine. [online] Statweb.stanford.edu. Available at: <https://statweb.stanford.edu/~jhf/ftp/trebst.pdf>.

Ghawi, R., and Pfeffer, J., 2019. Efficient Hyperparameter Tuning With Grid Search For Text Categorization Using Knn Approach With BM25 Similarity. [online] https://www.degruyter.com/. Available at: https://www.researchgate.net/publication/334597254_Efficient_Hyperparameter_Tuning_with_Grid_Search_for_Text_Categorization_using_kNN_Approach_with_BM25_Similarity.

Bergstra, J., and Bengio, Y., 2012. Random Search For Hyper-Parameter Optimization. [online] Jmlr.csail.mit.edu. Available at: http://jmlr.csail.mit.edu/papers/volume13/bergstra12a/bergstra12a.pdf.

Nerdy, F., 2020. What Is R Squared And Negative R Squared - Fairly Nerdy. [online] Fairly Nerdy. Available at: http://www.fairlynerdy.com/what-is-r-squared/#:~:text=It%20means%20you%20have%20no,worse%20than%20the%20mean%20value.

Sas.com. 2020. What Is Data Mining?. [online] Available at: https://www.sas.com/en_us/insights/analytics/data-mining.html.

Mitchell, T., 1997. Machine Learning. 1st ed. New York: McGraw Hill.

Sas.com. 2020. Machine Learning: What It Is And Why It Matters. [online] Available at: <https://www.sas.com/en_us/insights/analytics/machine-learning.html>.

Russell, S. and Norvig, P., 2009. Artificial Intelligence. 3rd ed. Upper Saddle River, New Jersey: Pearson Education, Inc., pp.695-727.

Rokach, L., and Maimon, O., 2005. Top-Down Induction of Decision Trees Classifiers—A Survey. IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews), 35(4), pp.476-487.

Schlesinger, S., 1979. Terminology for model credibility. SIMULATION, 32(3).

Deaton, M. L. (2006), "Simulation models, validation of", in Kotz, S.; et al. (eds.), Encyclopedia of Statistical Sciences, Wiley. [online] Available at: https://onlinelibrary-wiley-com.libproxy.tuni.fi/doi/10.1002/0471667196.ess2450.pub2#ess2450-bib-0009

Stuetzle, W., 2005. Cross-Validation. Encyclopedia of Statistics in Behavioral Science.

Stone, M., 1974. Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society: Series B (Methodological), pp.111-133.

Scikit-learn.org. 2020. 3.2. Tuning The Hyper-Parameters Of An Estimator — Scikit-Learn 0.22.2 Documentation. [online] Available at: <https://scikit-learn.org/stable/modules/grid_search.html>.

Claesen, M., and De Moor, B., 2015. Hyperparameter Search In Machine Learning. [online] arXiv.org. Available at: https://arxiv.org/abs/1502.02127.

Hsu, C., Chang, C., and Lin, C., 2003. A Practical Guide To Support Vector Classification. [online] Csie.ntu.edu.tw. Available at: <https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>.

Scikit-learn.org. 2020. 3.2.4.3.6. Sklearn.Ensemble.Gradientboostingregressor — Scikit-Learn 0.23.1 Documentation. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html.

Hastie, T., Tibshirani, R., and Friedman, J., 2009. The Elements Of Statistical Learning. 2nd ed. pp.364-369.