Monica Sidarta

Predicting Truck Cycle Time in Earthworks Using a Machine Learning Approach

TU Delft

bam

Predicting Truck Cycle Time in Earthworks Using a Machine Learning Approach

By

Monica Sidarta

in partial fulfilment of the requirements for the degree of

Master of Science
in Construction Management and Engineering

at the Delft University of Technology,
to be defended publicly on Thursday, July 8, 2021, at 08:30 AM.

Thesis committee:	Prof. dr. ir. L. A. Tavasszy	TU Delft
	Dr. M. Nogal Macho	TU Delft
	Dr. M. T. J. (Matthijs) Spaan	TU Delft
	Dr. Panchamy Krishnakumari	TU Delft
	Sreelatha Chunduri, Ph.D.	Royal BAM Group nv

Preface

The idea of this thesis is originally stemmed from my motivation to improve workers welfare in the construction industry and my interest in computational design. Based on my work experiences, inaccurate time prediction is a common problem affecting projects and workers. Workers work the hardest and being pushed to deliver the project if a project goes wrong. I also believe that digital and technology can improve construction management. Computational design has been a specific interest for me in which I did my bachelor thesis in Architecture. Fascinated by what digital and technology can do, I am encouraged to learn more, includes machine learning. Therefore, I am motivated to use it for my thesis research for improving time prediction accuracy.

I faced many challenges and surprises in the research process, such as the difficulty in obtaining data and reaching people. However, through the process and people I work with, I realized my knowledge and skill expanded. I also grow as a person, learn how to learn, and love what I do. I am grateful that I got supported from BAM, especially Digital Construction Department, and many people. Therefore, I would like to express my deepest gratitude.

I would like to thank my thesis committee: Prof. Tavasszy, Supervisor Matthijs Spaans, Supervisor Maria, Panchamy, and Sreelatha. My committee chair, Prof. Tavasszy, thank you for giving me constructive feedback and leading the meetings. Supervisor Matthijs Spaans, thank you for providing constructive feedback, which led my research in the right direction. Supervisor Panchamy, thank you for helping me understand the data and reminding me to connect my research with my knowledge domain and not define this thesis as my sole happiness. Supervisor Maria, I am honoured to know you and have you as my main supervisor. Thank you for your time, guidance, and encouragement from the asset management course until the end of my thesis. Sreelatha, I can't thank you enough for your time, ideas, suggestions, support, and for being such a good role model.

I would like to thank my family, especially my mom, who supports her daughter ambition to study. Thanks to Samuel for reminding me to arrive at the right destination, not only fly high. Thanks to Dee and Sanadhi for your willingness to discuss machine learning. Thanks to Vivian, Nindy, Chen, Marisol, Ita, and others for being a ray of sunshine on cloudy days and for the tremendous support.

Special thanks to BTS for cheering, motivating, and comforting me with your music and existence. Especially Kim Soekjin and Min Yoongi, I learn to run at my own pace and give my best. *Borahae!*

And the last, I would like to give all the honour to God who created, saved me through Jesus Christ, and gave me hope through the Holy Spirit. Great is Your faithfulness. Morning by morning, new mercies I see. *Soli Deo Gloria.*

I hope you enjoy your reading and are encouraged to give your best in a future where the next generation can enjoy.

Monica Sidarta
Delft, June 2021

Executive Summary

Background

The construction industry is known as one of the largest industries in the world. It is also forecasted to grow due to the increasing demand for housing and infrastructure. Earthworks will also increase because they are involved at the beginning of most construction projects. The largest activity in earthworks is truck movement or hauling to transport material. Thus, the planning process of earthworks involves time estimation of truck movement in transporting the material in one cycle, which is defined as truck cycle time (*see figure 1*).

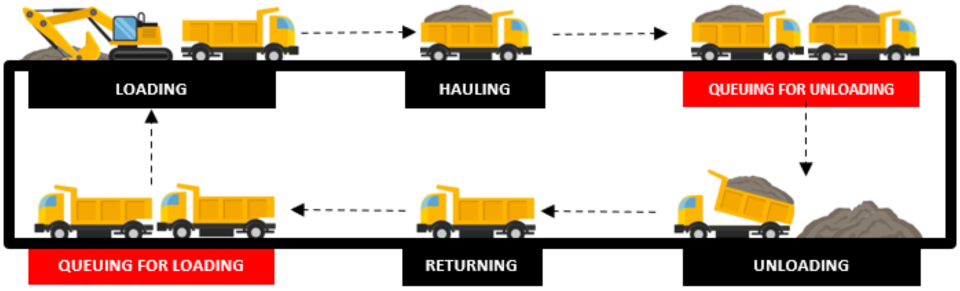


Figure 1. Truck cycle in earthworks

Experts currently calculate truck cycle time (TCT) by analyzing machinery specification, using a mathematical equation, or conducting a trial of a truck cycle on site before the project starts. However, the prediction is inaccurate because of subjectivity and human error, which easily affects the prediction. Inaccurate truck cycle time (TCT) prediction causes queuing of trucks for loading and unloading material. It causes a delay in completing the project on time and within budget due to a change in planning by adding additional equipment, machinery and human resources. It also negatively impacts the environment by increased fuel consumption, resulting in higher emissions. Hence, improving the accuracy of TCT is considered a critical element in increasing the construction industry's performance.

The machine learning (ML) approach is an approach to predict or make decisions. It uses historical data to build a model without being explicitly programmed. However, not much research has been conducted to develop a predictive model for TCT in earthworks using the ML approach. Thus, this research aims to utilize the historical data for improving the accuracy of TCT prediction in earthworks.

Research Scope

This research explores historical data gathered from projects of Royal BAM Group in the UK. The historical data were collected manually (manual data) and automatically collected using a machine (automated data). Both the data was explored and developed using regression techniques: Multi Linear Regression (MLR), Support Vector Regression (SVR), and Artificial Neural Network (ANN). However, this research did not explore the difference between historical data from BAM and other companies. Furthermore, this research did not consider the dependency between trucks in accurately predicting truck cycle time. It also did not include the government regulations on earthworks, such as operation time or the number of workers.

Research Methodology

The research results are obtained by answering the main research question: "*How can the historical data be utilized to improve the prediction accuracy of truck cycle time in earthworks?*". Three sub-questions are formulated to answer the main research question, as follows.

1. Which variables in the historical data should be included in the predictive model of truck cycle time in earthworks?
2. How to develop an accurate predictive model of truck cycle time using the machine learning approach?
3. What is the practical implication of using the predictive model of truck cycle time?

Each question was answered by following the scheme of research methodology, which is illustrated in Figure 2.

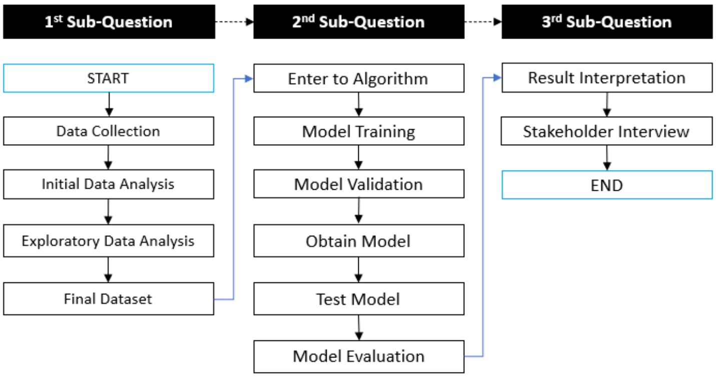


Figure 2. Research methodology

Findings

1. Data Exploration Result

Based on the literature review, factors of TCT are analyzed and used as the starting point to collect other data. This research also used weather data and combined it with the historical data. Each data is explored and analyzed through initial data analysis (IDA) and exploratory data analysis (EDA).

Data exploration shows that manual data has five variables and 430 data points, and automated data has seven variables and 589 data points. The availability of individual activity duration in automated data can predict TCT with three different scenarios.

- The first scenario is the accumulation of the load time model, haul time model, unload time model, return time model.
- The second scenario is the accumulation of the truck travel time, load, and unload time models.
- The third scenario is the truck cycle time model.

Hence, the variables are distinguished between input and output, shown in the following figure and used as TCT model input.

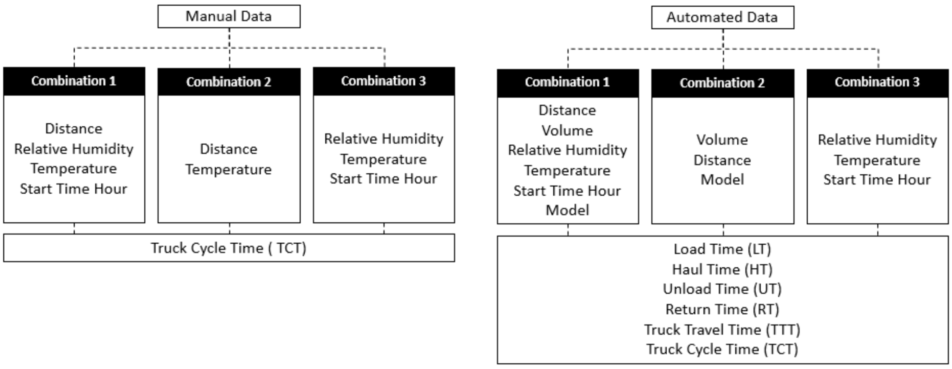


Figure 3. Input and Output

2. Modelling Result and Evaluation

The data was developed into predictive models, which has a low value of Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), and a high value of the coefficient of determination (R^2). Based on modelling outcome, the predictive models from manual data cannot be used, and automated data can be used. It was concluded that manual data quality and variance are insufficient to develop a robust predictive model. Hence, manual data was not used in predicting truck cycle time.

Table 1. Modelling Result

Output	Method	Input	Important Feature	Accuracy
LT	ANN	Distance, Volume, Relative Humidity, Temperature, Start Time Hour, Model	Distance	33%
HT	ANN	Distance, Volume, Relative Humidity, Temperature, Start Time Hour, Model	Distance	31%
UT	ANN	Volume, Distance, Model	Model	5.8%
RT	ANN	Volume, Distance, Model	Distance	78%
TTT	MLR	Distance, Volume, Relative Humidity, Temperature, Start Time Hour, Model	Distance	79%
TCT	ANN	Distance, Volume, Relative Humidity, Temperature, Start Time Hour, Model	Distance	56%

Table 1 shows the modelling result where ANN develops the most predictive models with feature combination one or two. It concluded that distance is an important feature for most predictive models. The comparison in predicting TCT from the test dataset concluded that the third scenario is the most accurate.

3. Practical Implication

The practical implication of the models was analyzed by interviewing two stakeholders - a project manager, and a logistics and operations analyst. The result from the interview gave insights into the current practices and the predic

tive models. The benefits of predictive models are calculated and analyzed by comparing the prediction result with the result from the traditional method. The test dataset represented two trucks in two days. Based on the calculation, scenarios are approximately 20% more accurate in predicting truck productivity. Scenarios can also decrease inefficient truck cycle time approximately five to six times from the traditional method. The reduction of inefficient truck cycle time has impacted the environment by reducing the fuel emissions and the number of human resources needed to complete the job. Based on the calculations and analysis, benefits for each stakeholder are concluded in Table 2.

Table 2. Benefits for stakeholders

Tangible Benefit	Intangible Benefit
Sub-Contractor	
• Reduce machinery emissions. • Reduce inefficient fuel and human resources cost	• Gain general contractor trust • Improved employee work satisfaction
General Contractor	
• Reduce machinery emissions. • Increase the accuracy of machinery productivity • Avoid contract penalty	• Better strategic plan to complete the project • A better decision in selecting sub-contractor/projects • Gain client trust • Safety • Improved employee work satisfaction
Supplier	
• Reduce the expense of overtime worker	• Increase employee wellbeing

Conclusion and Limitation

The historical data can be used for developing a predictive model using a machine learning approach. It can improve the prediction accuracy of TCT and offer benefits to the stakeholders. However, the predictive models have limitation for the input, which is shown in table 3. The input of the type of trucks for the predictive models is also limited: Caterpillar 745 and Volvo A45G. Also, the input of variable material for the predictive models is only Overburdened.

Table 3. Model limitation

Range	Distance (km)	Temperature (°C)	Relative Humidity (%)	Start Time Hour	Volume (m^3)
Min	0.6	8.9	49.3	7	1.14
Max	3.9	18.2	95.4	17	27.7

Recommendation

Recommendations were given to improve the predictive model. For instance, experts in the construction industry are suggested to raise awareness about the importance of data, improve earthmoving documentation, and improve the predictive models with better data collection methods using a machine learning approach.

Contents

List of Abbreviations

ANN	Artificial Neural Network
BP	Backpropagation
FP	Front Propagation
HL	Hauling Time
LT	Loading Time
ML	Machine Learning
MAE	Mean Absolute Error
MLR	Multiple Linear Regression
MSE	Mean Squared Error
PC	Principal Component
PCA	Principal Component Analysis
QTU	Queuing Time for Unloading
R^2	Coefficient of Determination
ReLU	Rectified Linear Activation Function
RMSE	Root Mean Squared Error
RT	Returning Time
SVR	Support Vector Regression
TCT	Truck Cycle Time
TTT	Truck Travel Time
UT	Unloading Time

List of Figures

List of Tables

1. Introduction

This chapter aims to introduce the base of this research. The general introduction will present a brief explanation of the importance of truck movement in the construction industry. Then, the problem from a practical and scientific perspective will be explained. The research objective, research questions, and the research methodology will also be presented.

1.1. Background

Truck movement or hauling is an important activity in the construction industry because it has a significant role in earthworks to carry the material from one location to another [1]. Figure 1 shows that hauling the material is the largest activity in earthworks with 45%. Moreover, earthworks is a major activity in large construction in terms of cost or time. It is needed to prepare the land for the upcoming construction activity, such as infrastructure or structure projects. Therefore, many stakeholders aim to improve truck movement in earthworks, for instance, by minimizing fuel consumption or increasing truck productivity. One of the essential parts of the truck movement is time prediction because it relates to trucks' productivity, the number of trucks and human resources, the type of machinery, and the maintenance treatment in the earthworks.

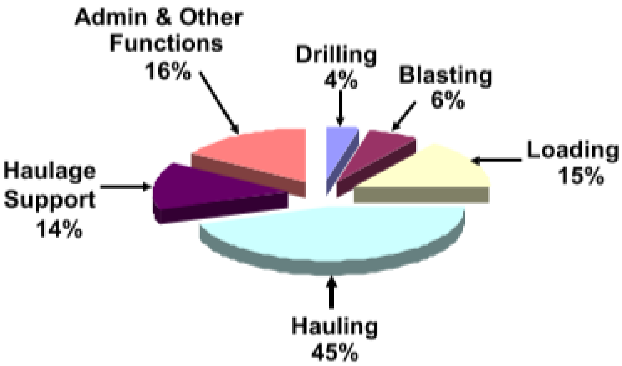


Figure 1. Activity in earthwork [1]

However, there is no sufficient method to accurately predict truck movement time due to the many variables that affect the calculation, such as the weather condition. It causes poor time prediction of truck movement and affects logistics management, which aims to manage the deliverance of the right material in the right quantity to the right place at the right time [2]. Poor logistics management might cause an inefficient strategy of managing machinery and human resources. Inaccurate time prediction may cause a delay in completing the project and require additional time to finish the project [3]. Additional time impacts project cost for adding more equipment and human resources. It also impacts the increase in fuel consumption and emissions from the machinery. According to many studies, the construction industry is responsible for up to 50% of climate change and severe negative impacts. It also caused 20% of worker fatalities, where logistics is one source of more than half of the fatalities [4]. Therefore, obtaining an accurate time prediction of truck movement is desirable in the earthworks because it can positively impact the strategy-making process, increase work efficiency, prevent workplace incidents and decrease environmental impacts [5].

Responding to the mentioned issue, this research aims to propose a method for improving the accuracy of time prediction of truck movement by considering several variables and assessing the practical implication of the proposed method in the stakeholders perspective in earthworks. This research was conducted as an internship at Royal BAM Nv.

1.2. Uncertainty in Estimating Truck Cycle Time

The construction industry is known as one of the largest industries in the world. It is also forecasted to grow at a compound annual growth rate (CAGR) of 4.8 from 2018 to 2023 due to the increasing demand for housing and infrastructure [6]. Earthworks are involved in most construction projects to prepare the area before constructing a new structure or infrastructure, such as roads and railways [7]. The earthworks activities can be considered a major part of a project because it costs high expense and duration. Hence, earthworks are considered a critical element of the overall project performance and affect the construction industry.

One of the vital parts in improving the earthworks is the planning process which ensures smooth and efficient project execution. The planning process of earthworks involves managing earthmoving machinery to transport the excavated material and construction material. The management of earthmoving machinery impacts the overall result and project cost due to the number of machinery, operators, and maintenance schedule. It is determined by calculating the earthmoving machinery productivity. One of the main earthmoving machinery is trucks because they have a relatively high speed and high flexibility in transporting materials [8].

The accuracy in calculating truck productivity is affected by time estimation. Time estimation of truck movement in transporting the material in one cycle is defined as truck cycle time (TCT). Inaccurate estimation of TCT may affect the project cost and project result whether the time is overestimated or underestimated. Overestimation may cause ineffective expense because the project pays for unnecessary machinery and resources [9]. Underestimation of TCT may cause poor project results due to machinery's low productivity and the number of resources. It may lead to the unavailability of the project on the stated completion date in the contract. The contractors have to pay the penalty and hire more human resources and equipment for completing the project. Therefore, an accurate time estimation of TCT is important for managing the resources in earthworks.

1.3. Method to Estimate Truck Cycle Time

The problem of estimating TCT relates to the effective prediction tool or method before the project is started. Currently, experts calculate TCT using the vehicle's specification from the machinery specification. Similar projects are often selected, analyzed, and adapted based on expert's experiences to predict the TCT. In addition, some contractors conduct a trial of a truck cycle on site before the project start. However, time prediction accuracy may be easily affected by subjectivity and human error [10]. As a result, more than half of the construction projects' deliveries are delivered in the wrong location and time [11].

Experts from research and practice backgrounds are trying to cope with human error in delivering the projects by utilizing information technology. Information technology is rapidly developed and used in recent years and becomes an essential aspect of the construction industry. Many data from previous construction projects have been documented and stored digitally. Sensors are also used for obtaining more detailed and accurate information. The development of the computer science field in terms of methods and hardware open many possibilities to develop new approaches to enhance industry performance.

In recent years, the machine learning (ML) approach has been used in many industries, including the construction industry, to predict or make decisions [12]. It uses historical data to build a model without being explicitly programmed [13]. However, not much research has been conducted to create a predictive model for TCT in earthworks using the ML approach. ML can be a good method for predicting TCT accurately in earthworks because of its capability to learn from historical data.

1.4. Research Objective

The research aims to improve the accuracy of TCT prediction by utilizing the historical data in earthworks. The output of this research will be a predictive model developed using a machine learning approach. The predictive model will be evaluated based on its accuracy and the practical implementation from the stakeholders perspective. The evaluation result of the model will be a valuable insight for potential opportunities and threats in using the predictive model in the future. In addition, it also may lead to a new perspective about the importance of historical data in improving the construction company performance.

1.5. Research Scope

The research has limitation due to the resources limitation and time constraint. The main limitation is related to the source of historical data and the research process. This research will only explore the historical data from projects of Royal BAM Group in the UK. The historical data were collected manually (manual data) and automatically collected using a machine (automated data). Also, this research will not explore the difference between historical data from BAM and other companies. Furthermore, this research will not consider the dependency between trucks in accurately predicting truck cycle time. It also will not include the government regulations on earthworks, such as operation time or the number of workers.

1.6. Research Questions

In responding to the mentioned problems and objective, the main research question is formulated, which is as follows.

How can the historical data be utilized to improve the prediction accuracy of the truck cycle time in earthworks?

The following sub-questions are formulated to answer the main research question in a structured manner.

1. Which variables in the historical data should be included in the predictive model of truck cycle time in earthworks?

The first sub-question objective is to identify variables that might affect TCT in earthworks. The process will require analyzing the main factors in TCT, which is investigated from previous literature. The result will lead to which data need to be collected, utilized and explored. Then, each variable in historical data will be explored and cleaned. Finally, the cleaned data will be used as the input for developing a predictive model.

2. How to develop an accurate predictive model of truck cycle time using the machine learning approach?

The second sub-question aims to develop an accurate predictive model of TCT by applying the ML approach. ML utilizes the cleaned data as the input and processes it for creating the prediction model. Different methods will be used to develop a predictive model of TCT. Each predictive model will be evaluated and analyzed.

3.What is the practical implication of using the predictive model of truck cycle time?

The third sub-question focused on investigating the practical implication of the predictive model. The selected stakeholders will be interviewed about the current practice and the predictive model. The implication of the predictive model in terms of cost and benefit will be evaluated. Moreover, the strategy to achieve an accurate predictive model of TCT will be presented to cope with the future challenge of practical implementation.

1.7. Research Methodology

The research methodology is formulated as the steps to answer research questions and achieve the research objective. Figure 2 shows the scheme of research methodology, which consists of three main steps, as indicates in each sub-questions. The following is a brief explanation of each step.

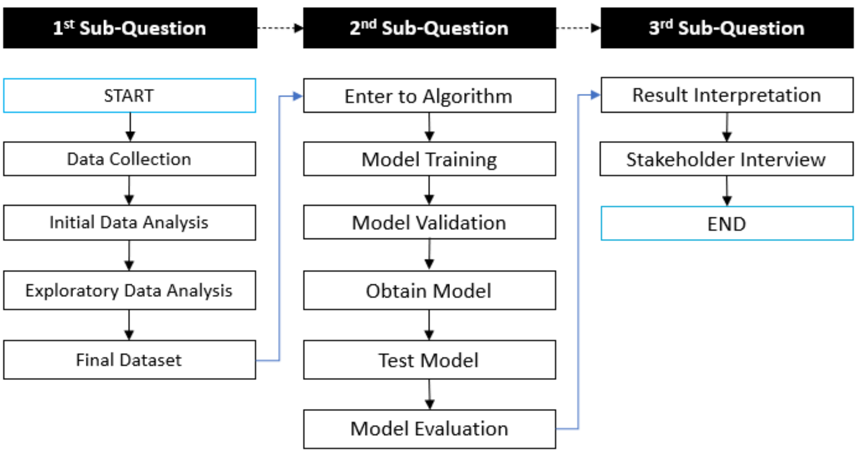


Figure 2. Research methodology

1. Investigating Related Variables

For answering the first sub-question, data will be collected based on the literature finding of the factors in TCT. This research uses a literature review to understand the problem gap and the solution space. The factors in TCT are used as the starting point to collect historical data effectively. Then, data preparation will be conducted where the historical data will be analyzed with initial data analysis (IDA) and exploratory data analysis (EDA). IDA is the process of data inspection to ensure data quality and minimize the risk of misleading results [14]. EDA is the later statistical analysis using data visualization methods, such as a scatter plot, Principal Component Analysis (PCA), and correlation matrix. Data preparation is the backbone of this research because the quality of a predictive model is dependent on the quality of data. Hence, it is important to understand the data by exploring and analyzing the data.

2. Developing Predictive Modelling

The machine learning approach is used for answering the second sub-question. This research will use the final dataset as ground truth and process it with a supervised approach from ML, where the machine learns to predict the outcome based on a training dataset. This process involves preparing the training dataset and test dataset, normalization, and hyperparameter tuning. Finally, methods for developing a predictive model will be selected based on the data analysis.

Each method will utilize a training dataset to learn and develop a predictive model. The predictive model is tested using a testing dataset for knowing its robustness. The test result will be evaluated and analyzed using performance metrics and compared with a different predictive model. The selected predictive model will be converted to the original unit and analyzed by denormalizing the result. The contribution of variable and the deviation value between prediction and actual value will be investigated using feature ablation.

3. Investigating Practical Implication

The practical implication of this research will be investigated from stakeholders perspectives by interviewing them to answer the third sub-question. The interview will be conducted with the selected stakeholders. The interview aims to gain stakeholders insight into the current practice, the advantages and disadvantages of the predictive model of TCT, interesting findings found in data preparation, and the future opportunities of ML in the construction industry. The result will be formulated as the strategy in the predictive model to improve the accuracy of TCT prediction.

1.8. Report Structure

This research methodology is structured to be the research report illustrated in the following figure.

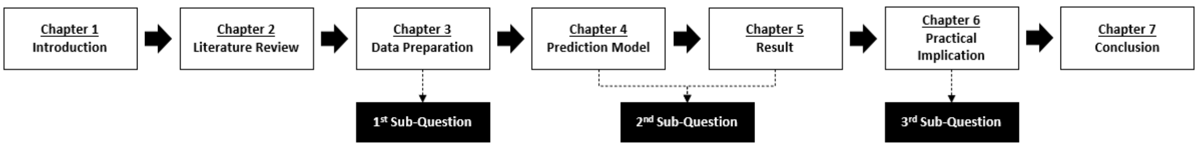


Figure 3. Report structure

According to Figure 3, this research is structured with seven chapters. The first chapter explains how this research is formulated by introducing the problem, objective, and how to address the problem by research questions and methodology. The second chapter aims to explain an extensive analysis of the literature finding of the problem gap, the main factors and the ML approach. The third chapter aims to answer the first sub-question through data preparation. Chapter four and five are dedicated to answering the second sub-question. In the fourth chapter, the final dataset is processed by algorithms to develop a predictive model. The predictive model will be evaluated and explained in the fifth chapter. Chapter six aims to answer the third sub-question by explaining the practical implication of the predictive model. In the end, Chapter 7 concludes this research and recommend future research.

2. Literature Review

The previous chapter introduced the overview of this research. This chapter aims to examine the relevant papers to investigate the previous research related to predicting TCT. Therefore, this chapter will investigate the earthmoving machinery works and compare the relevant papers. From these relevant papers, the main factors which affect the performance of earthmoving machinery will be analyzed. Data analysis and the machine learning approach are also explained in more detail.

2.1. Previous Research

Responding to the poor estimation of TCT, many experts tried to find the solution by proposing a new method to predict TCT accurately. Table 1 shows previous research which worked on improving the TCT prediction. Two main aspects are examined, such as the objective of the research and the method used.

Table 1. Previous research

Authors	Objective	Method
Plaistowe, R.H.A, et al. [16]	Calculate truck cycle time	Mathematic calculation
Peurifoy, R.L. et al. [8]	Calculate truck cycle time	Mathematic calculation
Curi et al. [17]	Evaluate the average fuel consumption and truck Cycle time	Effective flat haul (EFH)
Cervantes, E.G., et al. [9]	Generating accurate and reliable estimates of loaded truck travel times	Statistic simulation using data tonnes per gross operating hour
Sun, X. et al. [18]	Explore the result of TCT on the two different types of roads.	KNN, SVM, and RF

Before entering the digital era, experts use the manual calculation method to predict truck cycle time. Plastowe, R.H.A, et al. (1979) and Peurifoy, R.L., et al. (2018) proposed mathematic calculation for calculating Truck Cycle Time that can be done through manual calculation or without digital simulation.

Plastowe, R.H.A, et al. (1979) proposed the calculation of truck cycle time based on three major components, such as running, waiting, and loading. The research defined a constant in the TCT, which is from the calculation result of truck efficiency. The method proposed to sum up the running constant, accepted waiting constant, and loading time depending on the truck type. The result is proved to be reasonably simple and easy to be implemented in estimating TCT. However, the method requires the estimation of value for each component, and it does not eliminate the accuracy of TCT.

Peurifoy, R.L. et al. (2018) stated the method of predicting the truck cycle time using the mathematical calculation in his book. TCT is predicted by summing up four components: load time, haul time, dump time, and return time. Each component is calculated through a mathematical formula which is shown below, except the dump time. Dump time calculation depends on the type of truck and the condition of the dumping area. The average dump time on the favourable condition is 30 seconds and on the unfavourable condition is 90 seconds [8].

Equation 1. Load Time

$$Load\ Time = Number\ of\ bucket\ loads\ x\ Bucket\ cycle\ time$$

Equation 2. Haul Time

$$Haul\ time\ (min) = \frac{Haul\ Distance\ (ft)}{88\frac{fpm}{mph}\ x\ Haul\ Speed\ (mph)}$$

Equation 3. Return Time

$$Return\ time\ (min) = \frac{Return\ \ Distance\ (ft)}{88\frac{fpm}{mph}\ x\ Haul\ Speed\ (mph)}$$

This method provides a more detailed calculation for each component and less human assumption for predicting TCT. However, the variables required for calculating each component are hard to know in the planning phase. For instance, obtaining haul speed is difficult because it relates to other variables, such as the type of trucks, the amount of material, and the road condition. Hence, the method's accuracy is unknown because it relies on the experience of the expertise to predict the value for the variables.

Some experts propose a new method to estimate TCT accurately using record data with different approaches. For example, Curi et al. (2014) proposed a method to predict HT using the effective flat haul (EFH) parameter. EFH is defined as a calculated parameter that normalizes the elevation change of the route and the distance. This research applied the method toward two different trucks and two types of elevation change of haul routes to calculate the equipment's average cycle time. The research concludes several influence factors that can predict truck cycle time, such as material and site conditions. However, the accuracy of the prediction is unknown.

Cervantes, E.G. et al. (2018) used the record data of haul time, haul distance variability, and productivity performance indicator in tonnes per gross operating hour (TPGOH) and processed them using MATLAB to make a prediction model of haul time. This method's outcome is a curve that shows the relationship between the hauling time and the loaded haul distance. The result also shows an improvement in the accuracy, which compares with the EFH method. However, the method is only applied to predict the hauling time. Hence, the research suggests future research to develop a predictive model for RT.

Sun, X. et al. (2018) used ML with a classification method to predict TCT using the k-nearest neighbours (KNN), support vector machine (SVM), and random forest (RF) algorithms. The research develops a predictive model of TCT from two different routes: fixed-route and temporary. The research also includes the weather as the feature for the ML input. The research result shows that SVM and RF result is more accurate than the KNN. However, the research uses the classification algorithm for continuous value, which is not an effective way. And, the final accuracy is unknown because the error is not be normalized.

Previous research tried to use different methods and variables to obtain accurate prediction TCT. MATLAB and ML shows a significant improvement in accuracy and computational time by simulating the record data. The research which uses the ML approach gives a significant finding because of the variety of included variables and the complexity of the research.

2.2. Truck Cycle Time

This section aims to understand truck cycle time (TCT) before starting to develop a predictive model. TCT can be defined as the time estimation of truck movement in earthworks from a loading material location to an unloading material location. Figure 4 shows the illustration of a cycle of a truck transporting the material in earthworks. TCT consists of different activities in transporting the material, such as load time (LT), haul time (HT), queue time to unload material (QTU), unload time (UT), return time (RT), and queue time to load material (QTL). The following is a brief explanation for each activity.

Equation 4. Truck Cycle Time

$$TCT = LT + HT + QTU + UT + RT + QTL$$

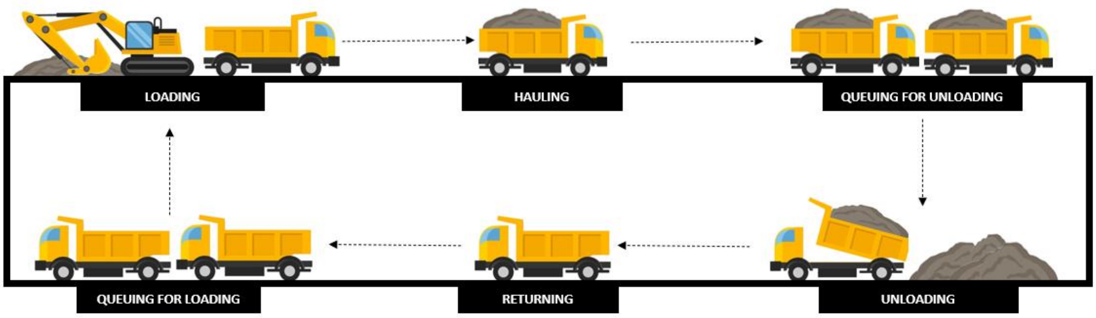


Figure 4. Truck cycle in earthworks

2.2.1. Load Time

Load time (LT) indicates the duration of an excavator to full in the truck bucket with the material. LT depends on the machinery combination between truck and excavator because the capacity of the excavator bucket affects the time needed to fill the truck bucket. An excavator with a large bucket capacity is often more expensive and consumes more fuel than an excavator with a smaller bucket capacity. Hence, the combination between excavator and trucks is important to load material quickly but at a low cost.

2.2.2. Haul Time

Haul time (HT) is the duration for a truck to transport the material to the dump location. The time needed depends on the speed of the vehicle to arrive at the unloading location. Hauling should concern the safe speed and road condition. Therefore, effective speed should be applied by the drivers.

2.2.3. Queue Time to Unload material

Queue time to unload material (QTU) is the duration for a truck to wait before its turn to unload the material. This activity is undesirable because it wastes fuel and human resources. The number of trucks and the size of the dump area impact QTU because the bigger size of the dump area will let trucks unload the material [18].

2.2.4. Unload Time

Unload time (UT) indicates the duration time for a truck to unload the material. This activity depends on the unloading area, which is usually crowded with support equipment, for instance, dozers [8]. Trucks will be more difficult to unload the material if the location is crowded because of limited space for a truck to manoeuvre and dump the material.

2.2.5. Return time

Return time (RT) is the time needed for a truck to arrive at the loading area from the unloading area. The main difference between RT and HT is the amount of material loaded in the truck bucket, where RT does not carry any material. In addition, maintenance of roads and equipment and operator behaviour also affect RT and HT [9].

2.2.6. Queue Time to Load Material

Queue time to load material (QTL) is the duration for a truck to wait before its turn is filled with material. QTL is different from QTU in the amount of material carried by the truck while waiting its turn.

QTU and QTL are affected by estimating the time needed for HT, RT, LT, and UT [19]. A truck that arrives later than the estimation time will cause the idle time of the excavator to fill the bucket or dozers to process the material. A truck that arrives earlier than the estimation time will cause the queuing for loading or unloading material. Therefore, the estimation of HT, RT, LT, and UT impacts machinery and human resources management.

2.3. Factors of TCT

The factor of TCT is examined to know which data should be collected and used for the ML process. TCT factors can be analyzed based on the element affected in the material movement activity and the literature about work efficiency in the material movement. Some prior research identified the significant factor in the productivity of machinery in earthworks. The factors are operation Practice, operating condition, and equipment [20]. Operation practice consists of the experience and habits of the driver when operating the machinery. The operating condition relates to the site condition and interaction between machines. Equipment defines as the technology in the used machinery.

Other research mentioned that two main factors affect TCT: relevant controllable and external factors [9]. The controllable factor is defined as road construction, safety guidelines, operator behaviour. And the external factors include weather condition and machinery repairment. Moreover, the report from Caterpillar mentioned that the production factors are machinery condition, operator skill, geological condition, and machine matching. Therefore, based on the previous research, it can be concluded that there are three main factors represented by operation practice, operating condition, and machinery condition.

Operation Practice, Operating Condition, and Machinery Condition

The factors of TCT is used as the guideline to collect and examine the data. The operation practice factor relates to the operator involvement in operating the machinery, including the training of the driver, driver behaviour, and experience. The operating condition factor relates to the site condition, including weather, road condition, type of soil, and material. And, the machinery condition factor relates to the machinery condition, which includes the machinery combination, number of the truck, type of truck. These factors contribute to the activity in earthworks which is illustrated in Figure 5.

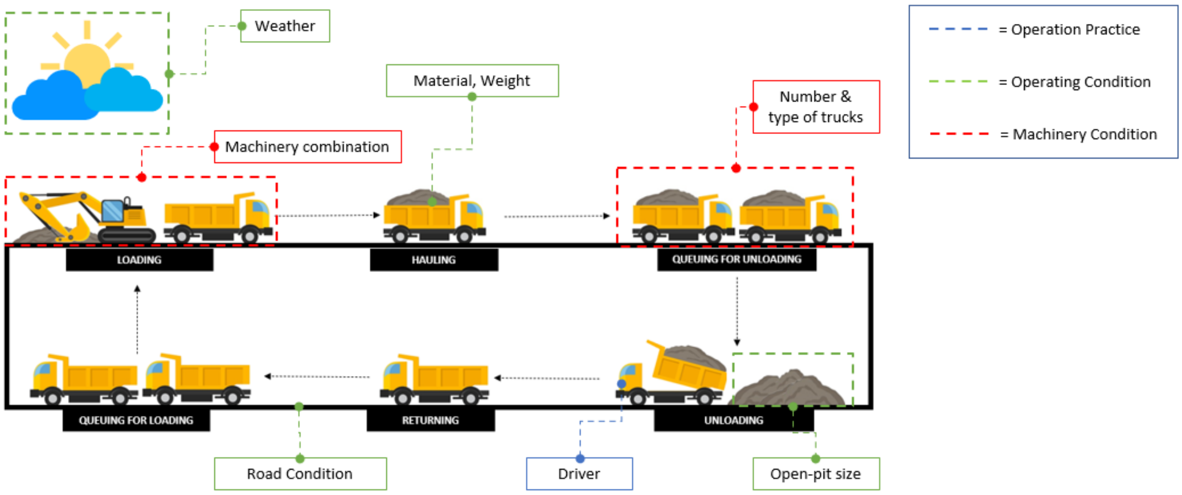


Figure 5. Factors of TCT

2.4. Machine Learning

Artificial Intelligence or AI has become more popular in recent years and has been applied in many industries, such as virtual assistant and self-driving cars. Artificial Intelligence is a built intelligence in a machine programmed to imitate humans' intelligence for doing particular tasks. It helps humans process many variables that are difficult or takes time to articulate into an output. Machine Learning (ML) is an approach that is commonly used in AI. The ML approach uses historical data to train the machine for predicting a certain output.

The type of learning in ML is divided into three main categories: supervised, unsupervised, and semi-supervised learning. Supervised learning is the type of learning required for human supervision to train the algorithm, for instance, to solve classification problems. Unsupervised learning is the type of learning that is not required human supervision. The result depends on the algorithm to learn and decide, for instance, to solve a clustering problem. Semi-supervised learning is the combination of supervised and unsupervised learning.

Along with the development of computer, ML algorithm can train more data and predict more complicated output using a neural network algorithm known as Deep Learning (DL). Deep Learning is based on multilayered neural networks that can learn from vast amounts of data. Figure 6 shows the illustration of the relation between AI, ML, and DL.

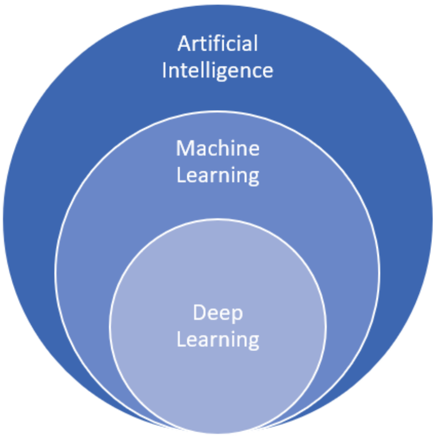


Figure 6. Artificial Intelligence, Machine Learning, and Deep Learning

The most important part of ML is to understand the target and choose the right model from ML to achieve it. In this research, the target is a predictive model of Truck Cycle Time (TCT). TCT is a continuous value that is suitable to be solved by a regression model. The regression model can be found in the traditional ML and DL.

2.5. Literature Gap

The previous section has explained that TCT is a pivotal element in the construction industry and has three main factors. Previous research shows that historical data can develop a predictive model by applying the ML approach. However, no prior research developed a predictive model of TCT using the historical dataset and ML approach. Therefore, this research will use the historical dataset to develop a predictive model of TCT by applying the ML approach.

3. Data Preparation

The previous chapter presented the machine learning (ML) approach and factors affecting TCT transporting the material. This chapter aims to explain the preparation of data to be a final dataset before entering the algorithm. Therefore, this chapter consists of data collection and analysis, consisting of initial data analysis (IDA) and exploratory data analysis (EDA). Then, the data is analyzed and cleaned by removing the error in the dataset to obtain a good quality of the dataset used as the final dataset.

3.1. Data Collection

This research tried to collect data that consist of variables in three main factors of TCT, such as the operation practice, the operating condition, and machinery condition. There are two main sources to obtain the data in this research. First, earthmoving machinery data is provided by BAM. The second source is Visual Crossing, which is a weather data service that provides weather data. Data from both sources consist of operating and machinery conditions, but the operation practice variable is unavailable. Therefore, this research only uses the available data and counts the unavailability of operation practice factor as the limitation of this research. The following is the explanation of how the data is collected from respective sources.

3.1.1. Earthmoving Machinery Data

The earthmoving machinery data consists of two record data from construction projects in the UK where BAM is involved as a contractor. The data are recorded in a different location, site condition and with a different method. The information in data related to the project name, location, and the serial number of vehicles are confidential. Hence, this research names the manual data entry with manual data and automated data entry with automated data. The engineer compiles both data into separate excel files. Therefore, both data will be understood and explored separately in this research.

3.1.2. Weather Data

Visual Crossing is a weather data service or weather API that can provide historical weather data and forecast weather data. This research collected the historical weather data by entering each cycle's start date and time from BAM data. Then, the weather data is added into earthmoving machinery data according to the date and time of data. Weather data consists of many variables which relate to weather. Table 2 shows the variables, type, and unit that is provided by generating the weather data.

Table 2. Variable, type, and unit from weather data

No	Variables	Type	Unit
1	Location Name	Object	N/A
2	Date time	Datetime	N/A
3	Maximum Temperature	Float	Celsius
4	Minimum Temperature	Float	Celsius
5	Temperature	Float	Celsius

6	Precipitation	Float	mm
7	Snow	Float	cm
8	Snow Depth	Float	cm
9	Wind Speed	Float	kph
10	Wind Direction	Integer	N/A
11	Wind Gust	Float	kph
12	Relative Humidity	Float	Percentage
13	Conditions	Object	N/A

3.2. Data Analysis

This research will analyze manual and automated data using two main steps: initial data analysis (IDA) and exploratory data analysis (EDA). IDA is an initial exploration of data by checking data quality, detecting and treating missing value, outliers, and other problem. Data quality check aims to understand the data process and trust the data as the ground truth for a predictive model. In data, sometimes there is a missing value which makes the algorithm cannot process the data. The missing data should be checked and treated based on the analyzed result. In addition, outliers that are often contained in the data need to be checked and treated. Outliers might come from human or machine error in documenting the data. Outliers might also give new information regarding the data is taken. Hence, outliers will not be eliminated directly.

EDA aims to maximize the understanding of data that includes the relationship between variables and feature selection. A pairs plot, correlation matrix, and PCA are used in this research and explained in the following sections.

Pairs Plot

The data distribution and relationship of a variable with other variables can be analyzed using a pairs plot. Figure 7 is a sample of a pairs plot that will be used in this research. It also shows regression lines for each relation and confidence interval illustrated by the shadow area around lines. Confidence interval is the range value of uncertainty for a certain parameter. The shadow area that wide indicates a high uncertainty to be able to get an accurate result. In addition, the direction of the shadow area is analyzed to uncertainty tendency based on the given data distribution.

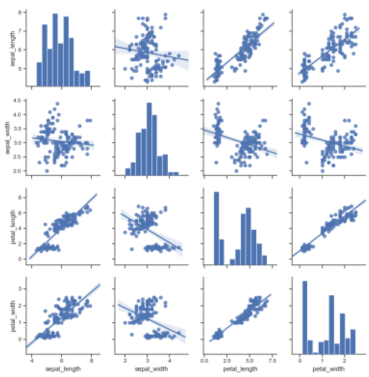


Figure 7. Illustration of a pairs plot

Correlation Matrix

The relation between all variables can be examined using correlation. The correlation value is between -1 and 1, representing a negative correlation for -1, a positive correlation for 1, and no correlation for 0. This research will use Spearman's correlation to assess the monotonic relationship, whether linear or not. Equation 5 shows the formula to find the spearman's correlation value which is indicated with ρ.

Equation 5. Spearman's Correlation Coefficient

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Where d_i refers to the difference between the ranks of each observation, n is the number of observations.

Principal Component Analysis

Principal Component Analysis (PCA) is a technique to emphasize variation and result in a strong pattern in data. PCA reduces dimensionality and makes the data easy to be interpreted with minimizing information loss. Figure 8 illustrates the comparison between data in original coordinates and principal component coordinates. The principal component axis indicates the direction that has high variance and more spread out. This research will use biplot to visualize principle components using python. It will show how strongly each characteristic influences a principal component.

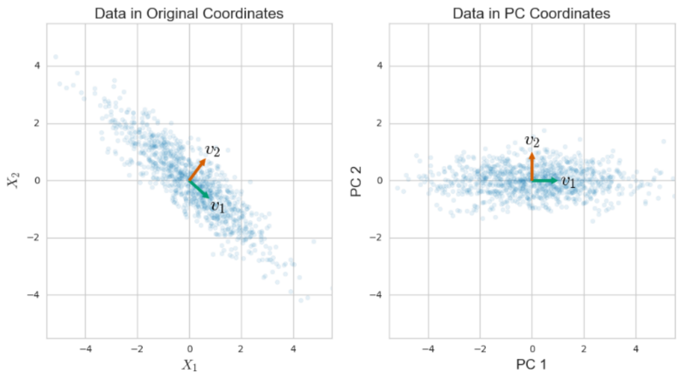


Figure 8. Illustration of how PCA work

However, each dataset contains a different range of value which can cause unequal calculation. Hence, normalization data can be used to transform the range between 0 and 1. The normalization function is as follows. Equation 6is the normalization equation, which is indicated by z_i. The input value is indicated by x_i. The maximum value and the minimum value are indicated by the $\max(x)$ and $\min(x)$, respectively.

Equation 6. Normalization

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

3.2.1. Manual Data

Manual data consists of 1500 data points or rows where each row represents one cycle and contains 14 variables not included in weather data yet. In combining manual data and weather data, the missing value in variable location makes the weather data not be generated. Hence, the data points in manual data decrease into 878 data points.

Table 3 consists of the list of variable, description, type, and unit of manual data. The data is taken by asking drivers to record their activity while driving the trucks to move the material in the project. Ten units of trucks operated in the four different months, between July and October in 2020. In combining manual data and weather data, the missing value in variable location makes the weather data not be generated. Hence, the data points in manual data decrease into 878 data points.

Table 3. Variables in manual data without weather data

No	Variables	Description	Type	Unit
1	Load Location	Location name for loading materials	Object	N/A
2	Load Latitude	Latitude coordinate of the loading location	Object	N/A
3	Load longitude	Longitude coordinate of the loading location	Object	N/A
4	Load Time	Date and time when loading materials	Date Time	N/A
5	Material	Type of Material	Object	N/A
6	Volume	Volume of loaded material	Integer	m^3
7	Weight	Weight of loaded material	Integer	Tonnes
8	Unload Location	Location name for unloading materials	Object	N/A
9	Unload Latitude	Latitude coordinate of the unloading location	Object	N/A
10	Unload Longitude	Longitude coordinate of the unloading location	Object	N/A
11	Unload Time	Date and time when unloading materials	Date Time	N/A
12	Total Cycle Time	Total duration for a truck in a cycle	Integer	seconds
13	Distance	Distance from loading to the unloading location	Float	meter
14	Vehicle	Vehicle identification number	Object	N/A

3.2.1.1. Initial Data Analysis (IDA)

Manual data quality depends on the guideline or standard in documenting the data to anticipate data errors, for instance, when drivers forgot to record the data. However, there is no clear guideline and uneasy access to the project documentation of the data collection process. For instance, the identity or the number of drivers is unknown because difficult to be traced back. Hence, this research needs to analyze the data deeply for checking the possibility of error in manual data.

Missing values in some variables are identified because of the unavailability of data on the requested date and time. For instance, snow data is not available because the requested date and time are in the summer. Variables containing many missing values are dropped from manual data to ensure the quality of data. Variable in manual data can be categorized into target and feature. Table 4 shows the variables in manual data that will be examined to detect and threat error in the data.

Table 4. Variables from manual data

Type	Factor	Variables
Target		Truck Cycle Time
Features	Operating Condition	Distance
		Temperature
		Relative Humidity
		Conditions

		Start Time Hour
	Machinery Condition	Volume
		Weight

Target: Total Cycle Time

The total cycle time (TCT) variable from manual data is recorded from the starting location or the loading location to the unloading location. According to the engineer, there is no queue time to unload the material and load the material. Equation 7 shows the calculation formula of TCT, and Figure 9 illustrates the recorded process in the data. The information about RT is unknown because of the lack of project documentation. Therefore, TCT from manual data is not included in RT and counted as the limitation of the predictive model.

Equation 7. Total cycle time in manual data

$$TCT = LT + HT + UT$$

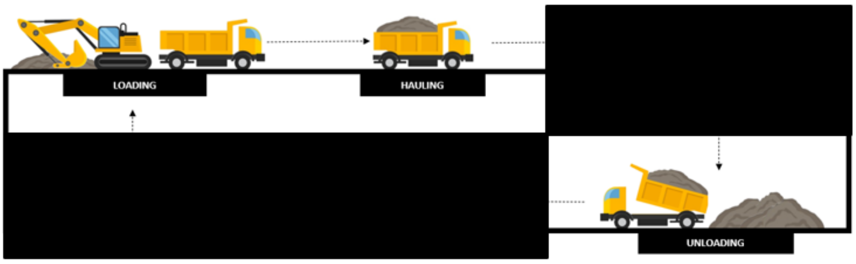


Figure 9. Illustration of TCT from manual data

Feature

Features of manual data are categorized into operating condition and machinery condition. Operating condition consists of distance, temperature, relative humidity, weather condition, and start time hour. Machinery condition consists of volume and weight. Each feature variable will be examined by visualizing the data points, and the problem will be detected and treated to obtain a good quality dataset.

Operating Condition: Distance

Distance variable from manual data indicates the distance from the loading location to the unloading location. The returning distance from the unloading location to the loading location is unknown. It does not have the same value as the hauling distance because the retuning path is different from the hauling path.

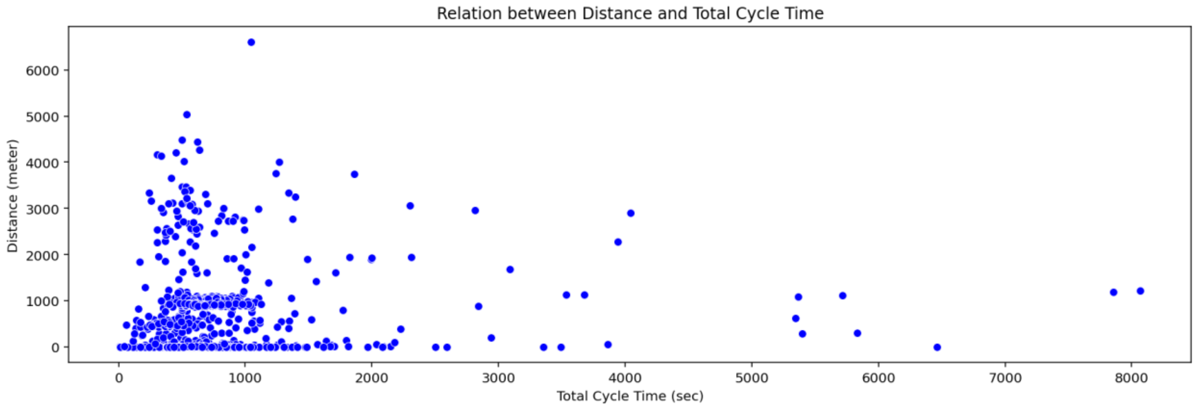


Figure 10. Scatter plot of distance from manual data

Figure 10 shows the relation of each data point based on the distance variable and TCT. The range of distance value is between 0 and 7 km. However, the data contains incorrect values since the distance value and TCT should not be zero. Zero value for distance and truck cycle time indicates a truck does not transport any materials. The incorrect value might be recorded mistakenly due to equipment limitation to record the data or human error. Therefore, data points that contain zero value for distance or TCT are eliminated as a part of data cleaning. As a result, the remaining data points is 536.

Operating Condition: Temperature and Relative Humidity

The cleaned data is analyzed in terms of temperature and relative humidity, aiming to detect any problem in the data. The range of temperature is between 6.5 and 31.5 degree Celsius from July to October 2020. Figure 11 shows the temperature value of manual data each month where the highest median value in August and the lowest mean value in October. Figure 12 shows overall temperature value is mostly distributed between 0 and 1000 seconds. Some outliers are detected in August, but they will remain because it helps consider the extreme temperature in the predictive model.

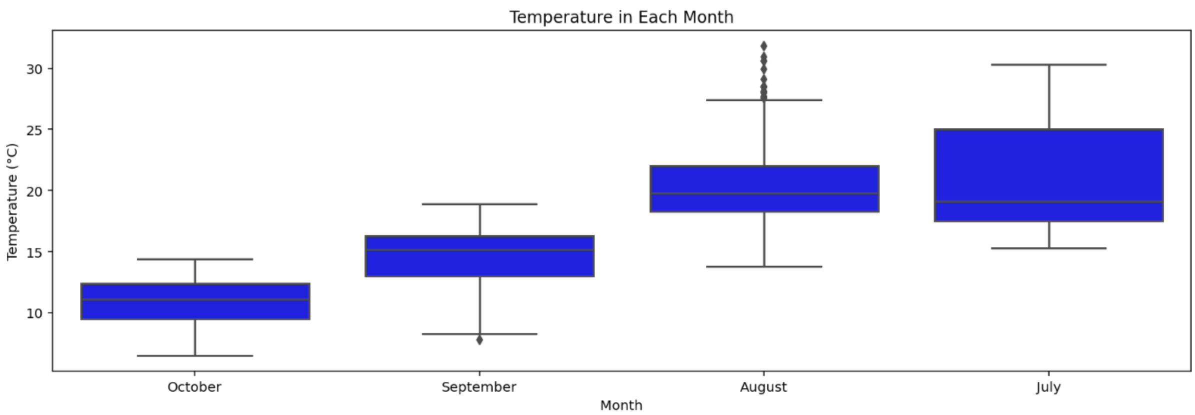


Figure 11. Boxplot of temperature for each month

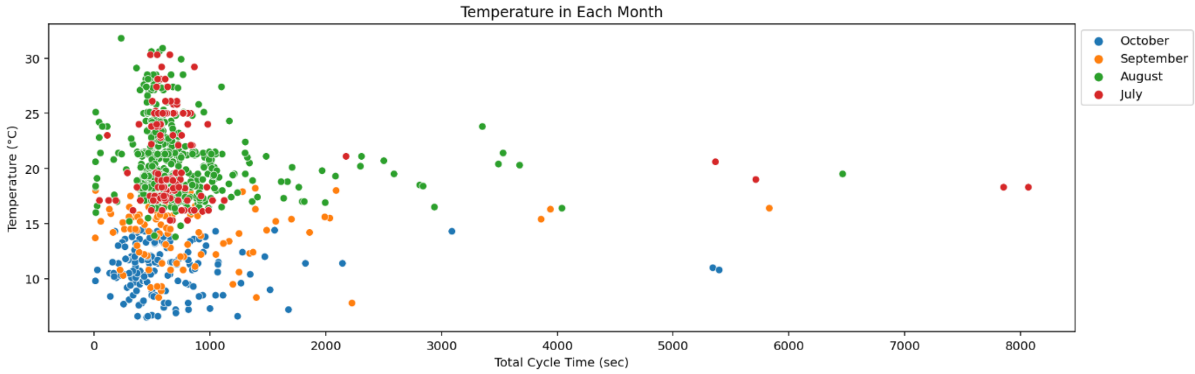


Figure 12. Scatter plot of temperature from manual data

The relative humidity from manual data is between 37.2 and 100 percent, with the same month range. Figure 13 shows the temperature value of manual data each month where the highest mean value in August and the lowest mean value in July. Figure 14 shows overall temperature value is mostly distributed between 0 and 1000 seconds. Although the range of total cycle time is similar between temperature and relative humidity, their value is not similar based on the month. The outliers in relative humidity will not be eliminated because it will help to consider unusual relative humidity for the predictive model.

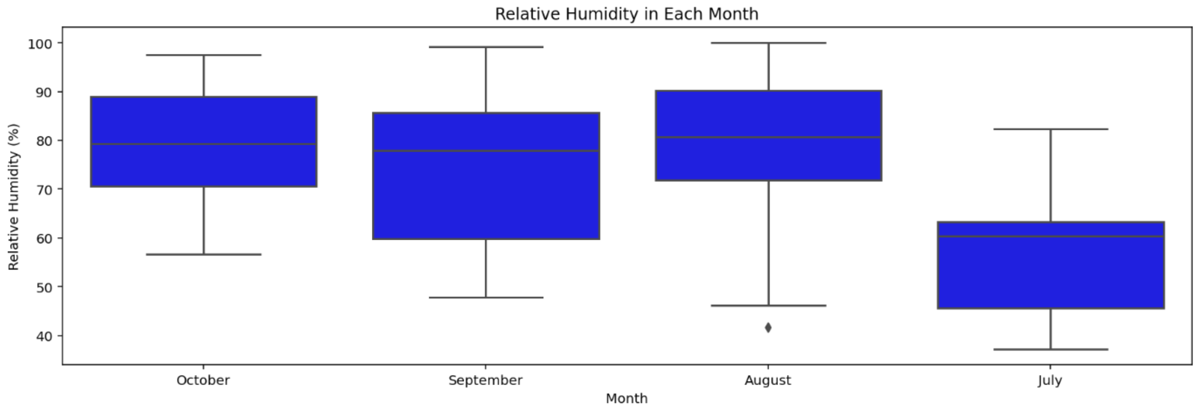


Figure 13. Boxplot of relative humidity in each month

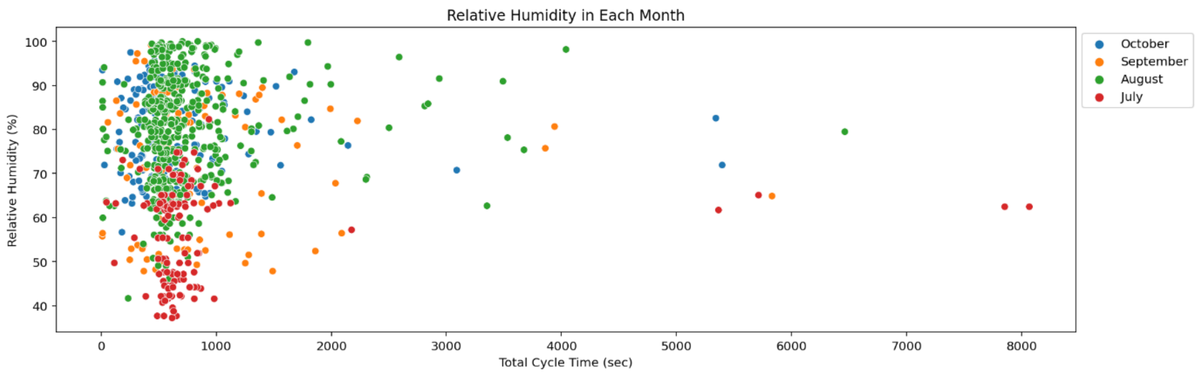


Figure 14. Scatter plot of relative humidity from manual data

Operating Condition: Weather Condition

Condition or weather condition refers to the sky condition such as cloudy and rain. Figure 15 shows the scatter plot of the condition in manual data. Clear is the only weather condition when the data is taken. Clear refers to no cloud in the sky and not rain. It indicates that the data is taken when the weather is good for moving the material. However, this information cannot be used because the machine will learn only one weather condition pattern. Thus, the weather condition will be eliminated from manual data and count as the limitation of the predictive model.

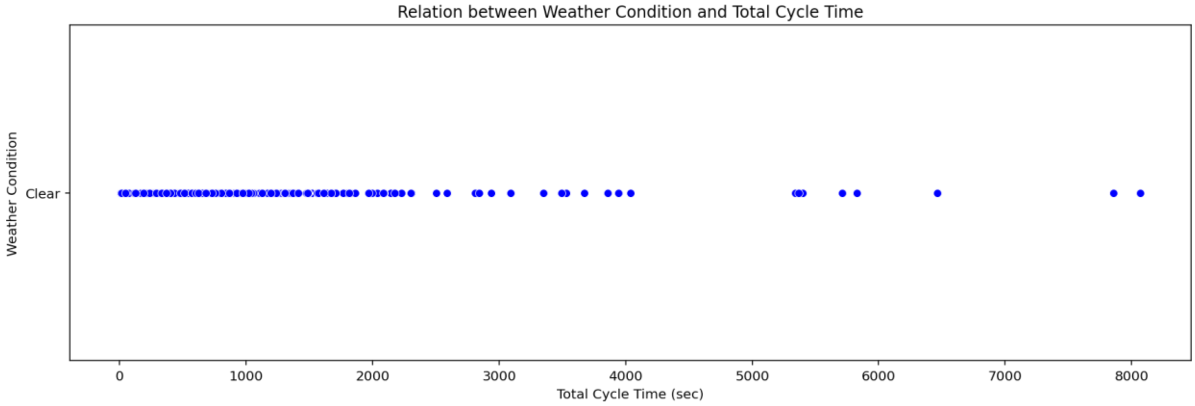


Figure 15. Scatter plot of condition from manual data

Operating Condition: Start Time Hour

Figure 16 shows a box plot to examine the data based on the truck's starting time transporting the material. Trucks are operated between 7 AM and 5 PM, where the highest truck cycle was started at 2 PM. Some outliers are detected and kept in the boxplots because there is no particular TCT at a certain time.

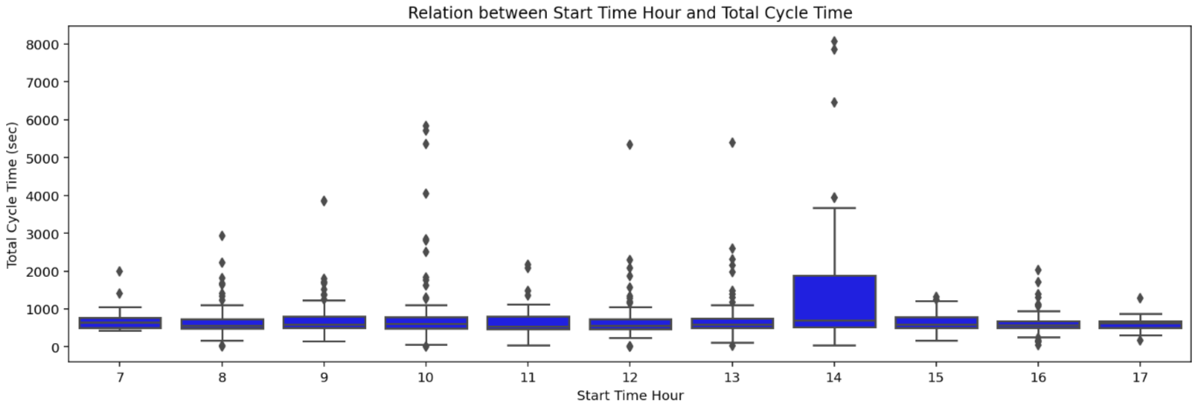


Figure 16. Box plot of start time hour from manual data

Machinery Condition: Weight and Volume

The earthmoving machinery that is recorded in manual data is an eight-wheeler tipper which shows in Figure 17. It is common to use this type of truck to move material on asphalt road. The weight capacity and volume capacities are 20 Tonnes and 15 m³, respectively. This information will be used to examine weight and volume data in manual data.

Figure 17. Eight Wheel Tipper

Data points contain zero as the value of volume and weight of the material that trucks are transported. Since TCT in data is indicated as a cycle transporting material from the loading to unloading area, it needs to be removed. The elimination of the zero value causes the reduction of the data point, which is 430.

Figure 18 and Figure 19 show the scatter plot of volume and weight of material in project A, respectively. The error data is detected because the volume value is 8 and 30 m³, and the weight value is recorded for 0, 18, and 20 tons for all data point. Those values are unreliable because weight value is usually continuous to value, and difficult to maintain the same value for each cycle. The error might result from the manual method of data collection and equipment limitation to scale the volume and weight accurately. Therefore, volume and weight are dropped from manual data.

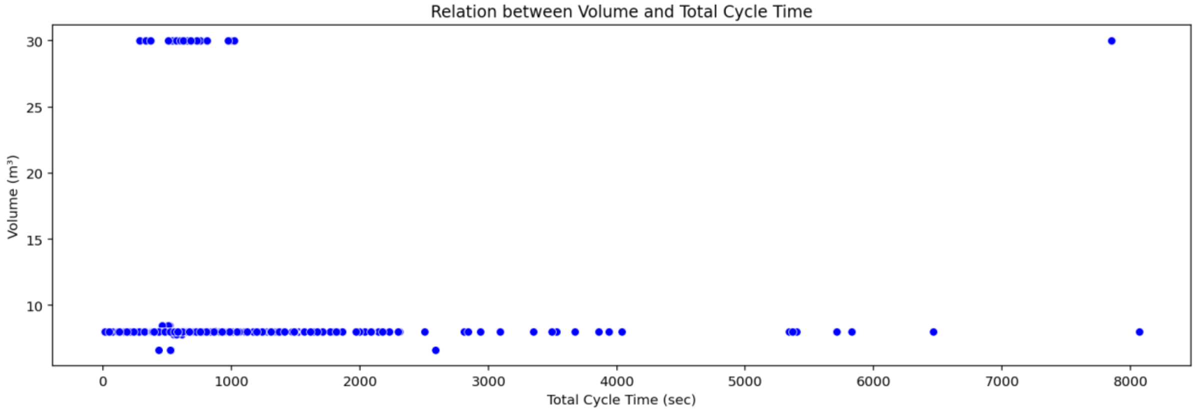


Figure 18. Scatter plot of volume from manual data

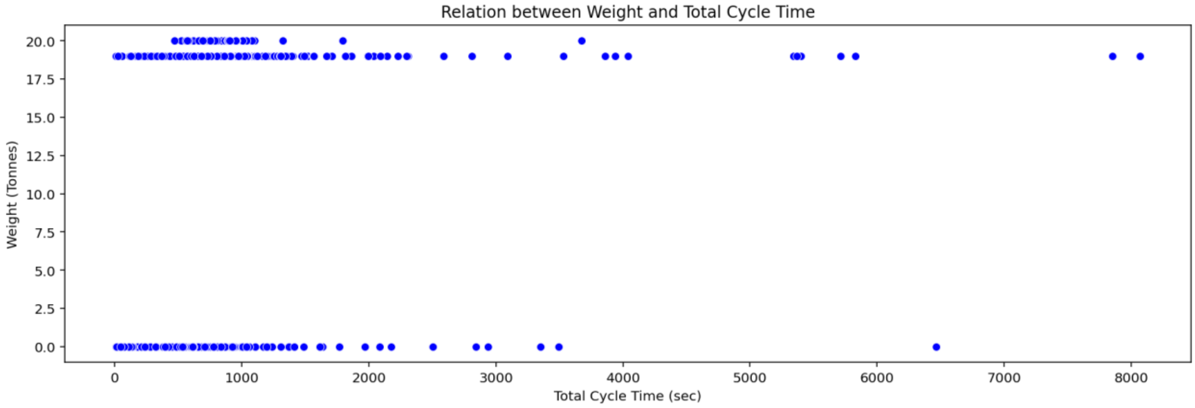


Figure 19. Scatter plot of weight from manual data

3.2.1.2. Exploratory Data Analysis (EDA)

Manual data is analyzed based on the relation between variables to understand patterns within the data. Figure 20 shows the pairs plot of the manual data, indicating the relation between variables and regression line. The last row of the plot shows the data distribution between each variable and truck cycle time without return time. The data distribution is not spread out equally, and most data is gathered between 0 and 1500 seconds. The lack of data above 1500 seconds causes the uncertainty of result from the regression line. The shadow area is located above and under the regression line, indicating that regression underestimates or overestimates time. However, suppose the regression and the shadow line is drawn farther. In that case, the regression line and the confidence interval between variable temperature and total cycle time will go upright. It indicates that the higher value of temperature will cause more time for a truck to transport material.

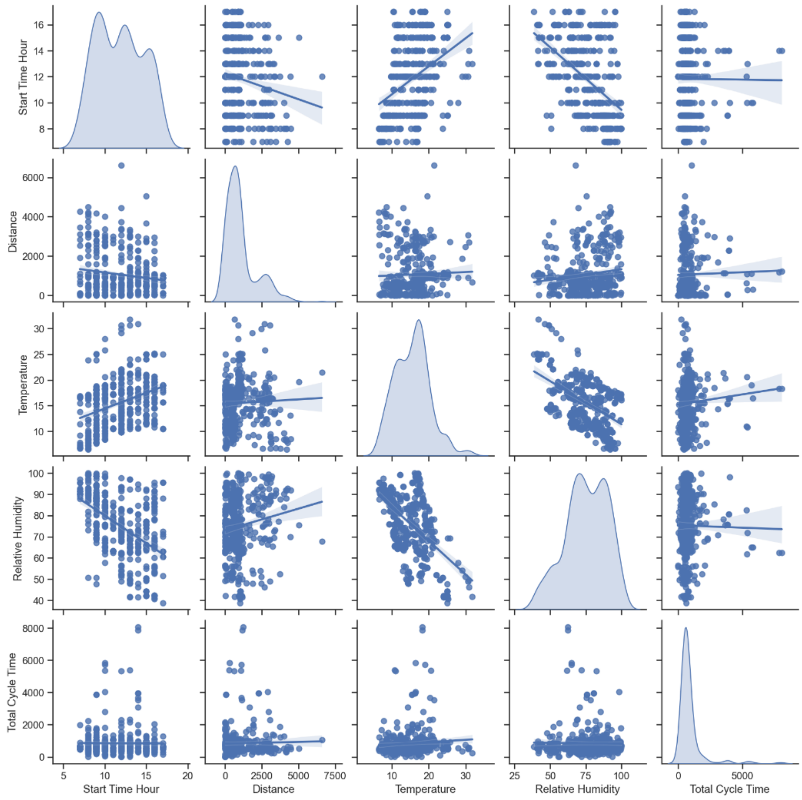


Figure 20. Pairs plot of the manual data

The pairs plot's result can be analyzed further with Figure 21, which shows the manual data's correlation matrix, which contains the correlation value. The correlation value between temperature and total cycle time is positively correlated and higher than the correlation between other variables and truck cycle time. The correlation values are positive except the correlation value between variable start time hour and truck cycle time. The correlation matrix helps to analyze the value, which is difficult to be captured by the pairs plot. It shows there is no high correlation value between variables which indicated the pattern of variables is different. Hence, variables can give a contribution and can be included to develop a predictive model.

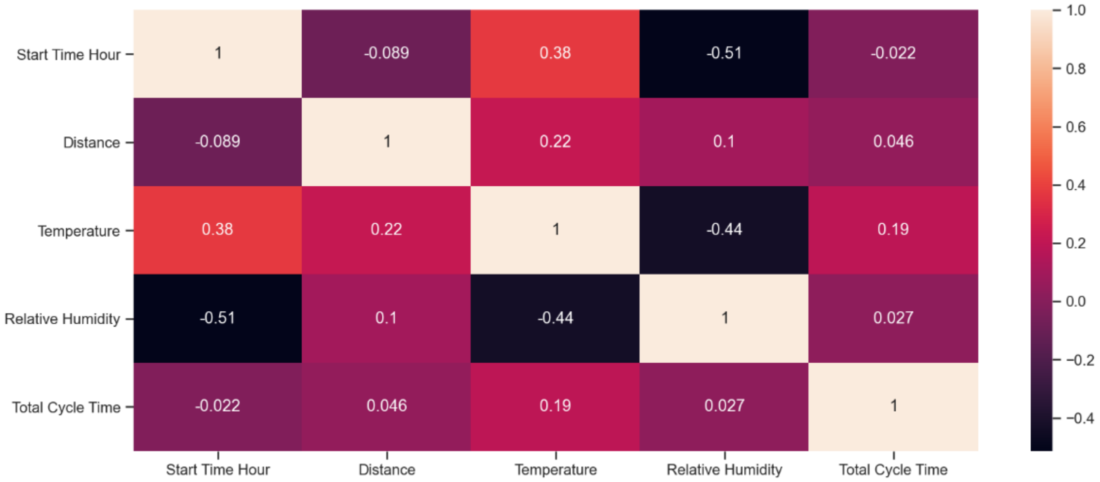


Figure 21. Correlation matrix manual data

Figure 22 shows the principal component analysis (PCA) between variables respectively from the manual data. It shows that both principal components (PCs) contribute approximately 75% of the total variation in the dataset. PC 1 explains 48,72%, and PC 2 explains 25.68% of the variation in the manual data. The difference between PCs depends on the influent of variables on the PCs. Distance gives more influence to principal component (PC) 2 than 1. Relative humidity and start time hour give more influence toward PC 1 than 2. The temperature has a similar influence on both PCs. The difference between PCs in capturing the variation can base model input in developing a predictive model.

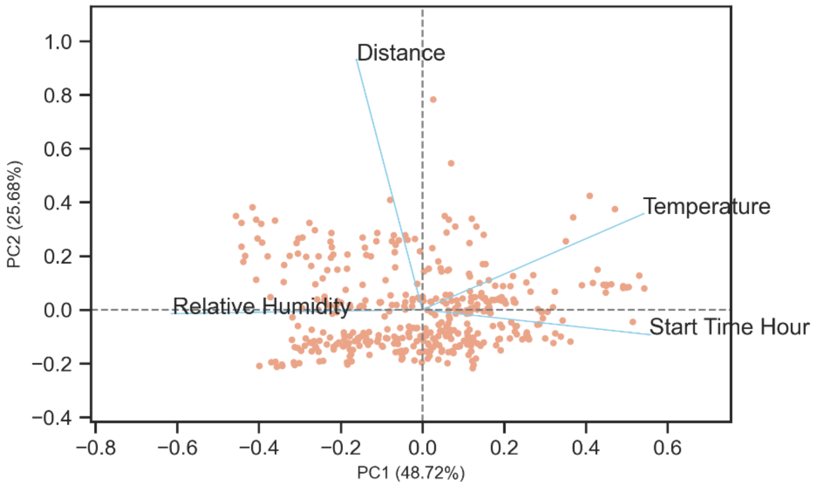


Figure 22. PCA manual data

3.2.2.Automated data

Automated data consists of 669 data points or rows which each data point consists of 21 variables not included in weather data.

Table 5 shows the detailed information of each variable, the variable description, type, and unit in the automated data. The automated data is recorded using an application connected with a sensor and earthmoving machinery to document the machinery activity. The application is provided by Caterpillar and not widely known yet because it is a new application. Automated data consists of data from 16 to September 17 2019, in project B.

Table 5. Automated data

No	Variables	Description	Type	Unit
1	Cycle Start Time	Start time of a truck cycle	Date Time	N/A
2	Cycle End Time	End time of a truck cycle	Date Time	N/A
3	Material	Type of Material	Object	N/A
4	Weight	Weight of loaded material	Integer	Tonnes
5	Volume	The volume of loaded material	Integer	m^3
6	Source Location Latitude	Latitude coordinate of the loading location	Object	N/A
7	Source Location Longitude	Longitude coordinate of the loading location	Object	N/A
8	Destination Location Latitude	Latitude coordinate of the unloading location	Object	N/A
9	Destination Location Longitude	Longitude coordinate of the unloading location	Object	N/A
10	Total Cycle Duration	Total duration for a truck in a cycle	Integer	seconds
11	Total Cycle Distance	Distance from loading to the unloading location	Float	meter

12	Start Location Name	Location name for loading materials	Object	N/A
13	End Location Name	Location name for unloading materials	Object	N/A
14	Load Time	Duration for loading materials into a truck	Integer	seconds
15	Haul Time	Duration for transporting materials	Integer	seconds
16	Loaded Stopped Time	Queueing time for unloading materials (QTU)	Integer	seconds
17	Return Time	Duration for picking materials	Integer	seconds
18	Empty Stopped Time	Queueing time for loading materials (QTL)	Integer	seconds
19	Unload Time	Duration for loading materials from a truck	Integer	seconds
20	Total Cycle Fuel Liter	Total fuel consumption for a cycle	Float	Liter
21	Loader Serial Number	Vehicle identification number	Object	N/A

3.2.2.1. Initial Data Analysis (IDA)

After adding the weather data into automated data, variables in automated data are categorized into target and feature variables. However, automated data doesn't provide the information about the latitude and longitude fully. Hence, the data points decrease from 669 to 589 because the data points which contain missing value are eliminated. Variables in automated data are categorized into target and feature. Table 6 shows the variables in automated data that will be examined to detect and threat error in the data.

Table 6. Target and Features from Automated data

Type	Factor	Variables
Target		Total Cycle Time (TCT)
Features	Operating Condition	Distance
		Temperature
		Relative Humidity
		Conditions
		Start Time Hour
	Machinery Condition	Model
		Volume
		Weight

Target: Truck Cycle Time (TCT)

Automated data consists of the record time of each activity in moving the material, such as Total cycle duration, LT, HT, QTU, RT, QTL, and UT. Total cycle duration is the outcome of the application that records a truck's duration in one cycle. In theory, the total cycle duration value should be as equal as finish time minus start time. However, the data shows that the Total Cycle duration value is bigger than the deviation time of the finish and start times. Furthermore, the total sum of LT, HT, QTU, RT, QTL, and UT equals the deviation value. Therefore, Total Cycle Duration is not reliable to be used as the target for this research.

QTU and QTL depend on the number of trucks operated in a day, the number of excavators to fill the truck bucket, and the size of the unloading material site. Automated data does not have any data about the number of excavators and the size of the site. And automated data only provides the total number of trucks in two days, which is insufficient to analyze QTU and QTL. Therefore, only LT, HL, RT, and UT are reliable for ML because of the limited information about QTU and QTL. The accumulation of individual activity is defined as total cycle time, which is illustrated in Figure 23

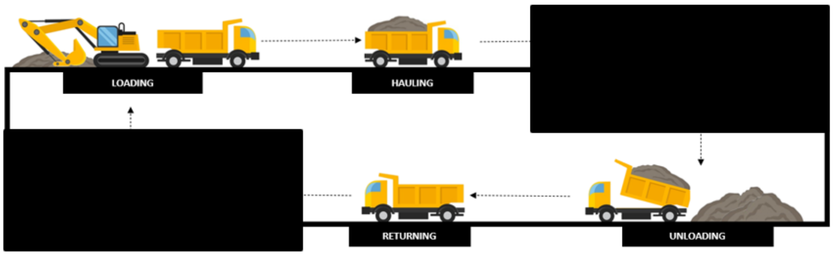


Figure 23. Illustration of TCT

Features

Distance, temperature, relative humidity, weather conditions, and start time are categorized in operating conditions. Model, volume, and weight are categorized into machinery condition automated data. Feature in automated data is examined by visualizing each feature with various plot types.

Operating Condition: Distance

Distance data in automated data is the accumulation value of hauling distance and returning distance. The distance for each activity is unknown due to the application is not designed to track each distance. Figure 24 shows no zero value for distance and TCT, which indicates each data point recorded material moving.

The range is between 0.6 km and 3.9 km, resulting from a different path, although the loading and unloading location are the same. It also shows that most trucks spend between 400 to 750 seconds for a distance between 1.2 and 2.0 km.

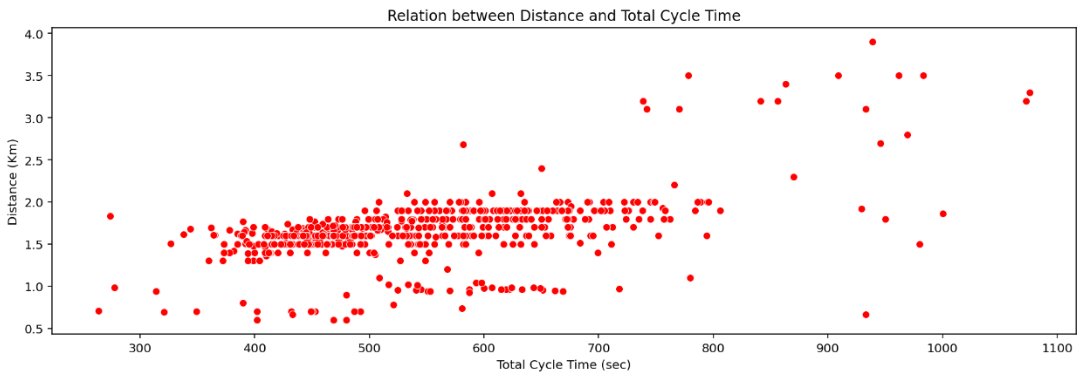


Figure 24. Scatter Plot of Distance from Automated data

Operating Condition: Temperature and Relative Humidity

Figure 25 and Figure 26 show that the temperature value is between 8.9 and 18.2 degree Celsius and the relative humidity value is between 49.3 and 95.4 percentage, respectively. Trucks that transported the material are mostly around 16 degrees Celsius and 50 per cent or 85 per cent of relative humidity.

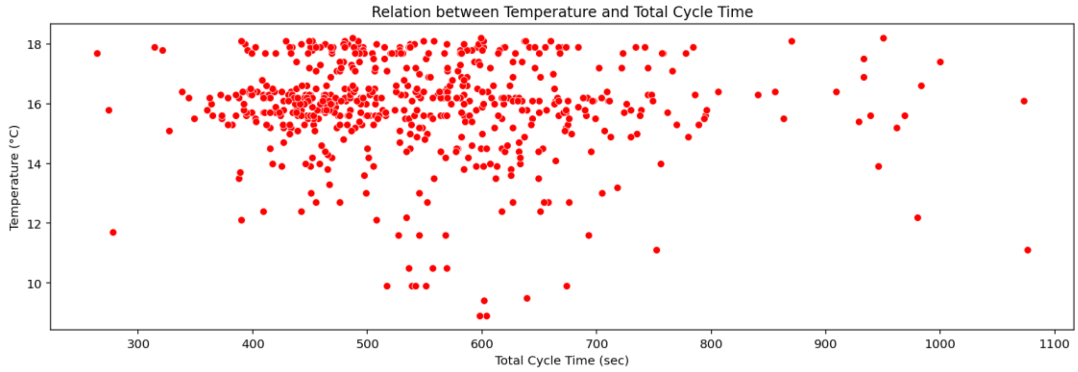


Figure 25. Scatter Plot of Temperature from Automated data

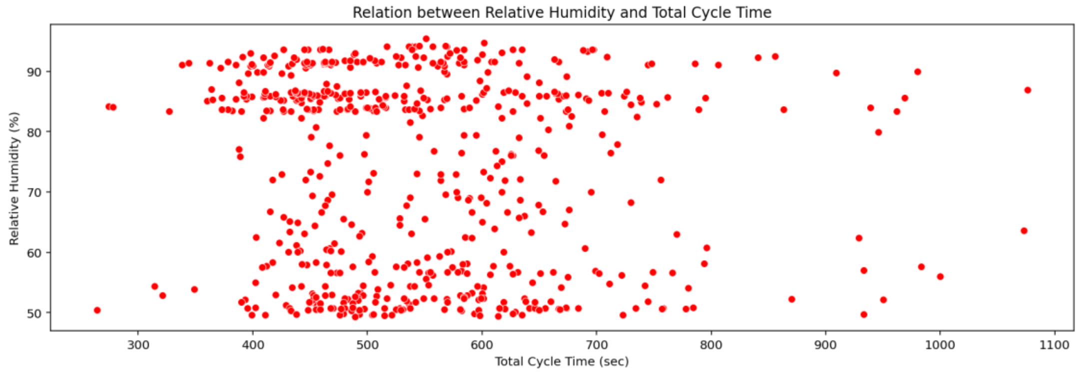


Figure 26. Scatter Plot of Relative Humidity from Automated data

Figure 27 and Figure 28 shows the box plot of temperature and relative humidity based on the starting time of a truck operation to move the material. Both box plots show a few outliers, but they might indicate extreme temperature or relative humidity in the future.

The lowest mean value of temperature and relative humidity is when the truck operated at 7 AM and 5 PM, respectively. Moreover, the widest range of relative humidity is at 5 PM. This finding shows that the truck cycle time is not highly impacted by temperature and relative humidity compared to the previous finding.

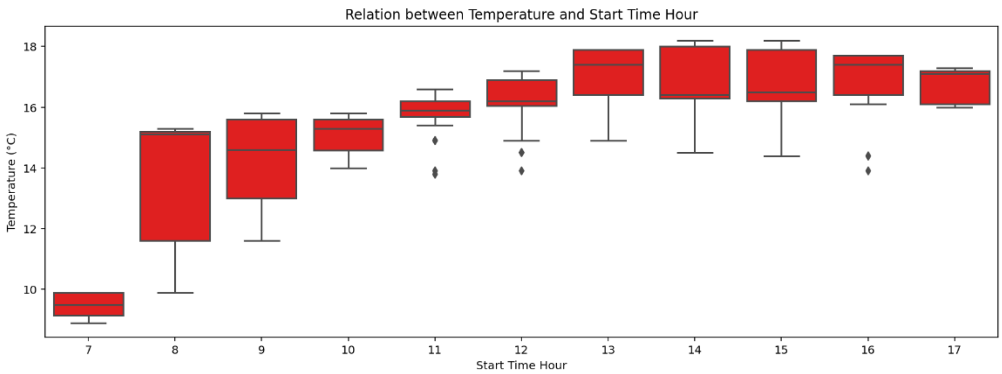


Figure 27. Box Plot of Temperature based on Start Time Hour from Automated data

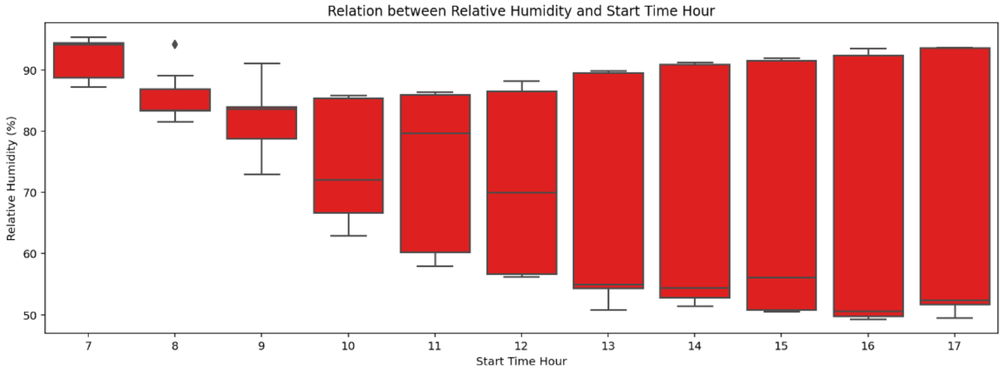


Figure 28. Box Plot of Relative Humidity based on Start Time Hour from Automated data

Operating Condition: Weather Condition

Figure 29 shows the weather condition when the data consists of clear, partially cloudy, rain, overcast, rain and overcast, and rain and partially cloudy. It shows significant unbalanced data between the type of condition where clear has the biggest data point. It impacts machine learning to learn and create the prediction model because it will not learn enough about weather conditions from low data. Some methods can be used for solving the unbalanced data issue. Still, it requires enough data to generate synthetic data or eliminate the condition with low data points. However, the amount of data point is not big, and it will decrease the potential of ML to learn by eliminating the type of condition. Therefore, the condition is eliminated from the feature and count as the limitation of this research.

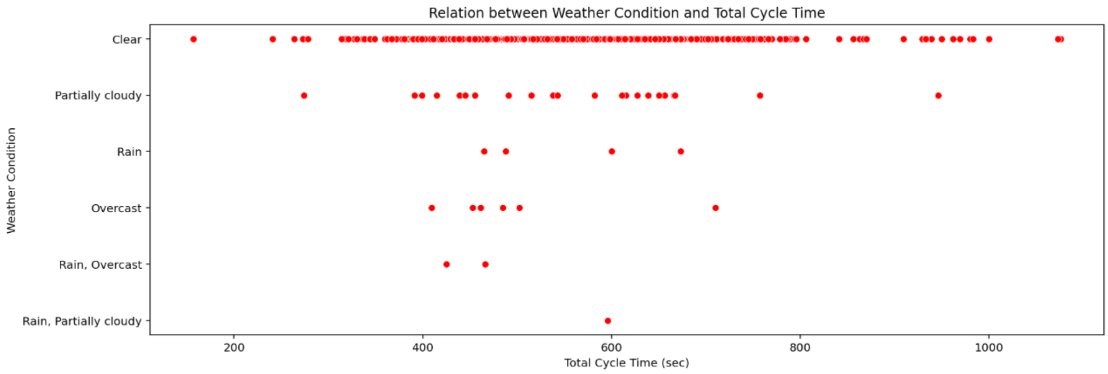


Figure 29. Scatter Plot of Condition from Automated data

Operating Condition: Start Time Hour

The Start Time Hour variable is an extracted result of the hour from the Cycle Start Time variable. Starting time when the trucks are operated are between 7 AM and 5 PM. Figure 30 shows the highest mean value for a truck finishing one cycle when the truck starts to operate at 7 AM. It shows that the truck number that starts to operate at 7 AM is lower and might cause slower operation behaviour to transport the material. It also shows that the range of value at 7 AM is lower than other start time hour.

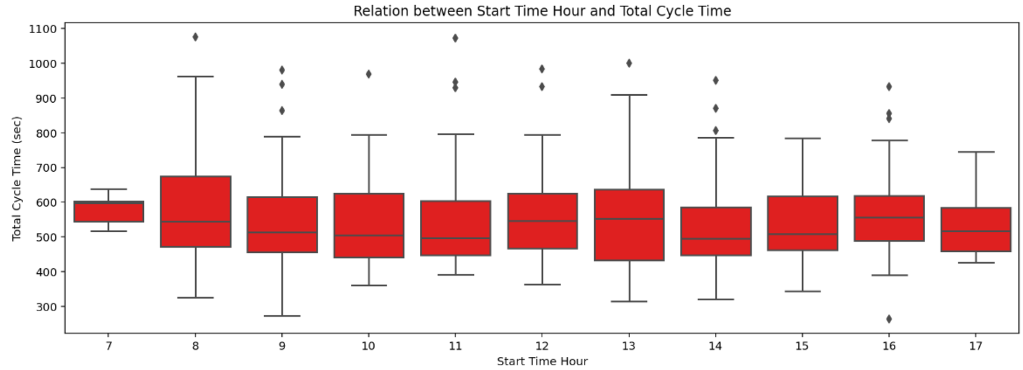


Figure 30. Box Plot of Start Time Hour from Automated data

Machinery Condition: Model

Automated data provides data about the model of the truck that is used to transport the material. The type of truck is Articulated Dump Truck (ADT) which is commonly used for moving the material in earthworks due to the capability to move in difficult soil conditions [8]. There are two different models of ADT, which are Caterpillar 745 and Volvo A45G. Both have the same weight and volume capacity, which are 45.3 Tonnes and 25 m³, respectively.

Figure 31. Type of Trucks in Automated data: Caterpillar 745 and Volvo A45G

The data need to be modified from an object into binary value so the machine can process it. Hence, Caterpillar 745 is replaced with 0, and Volvo A45G is replaced with 1. Figure 32 shows that the mean value of TCT of model 0 is higher than model 1. Model 1 is faster than model 0 to transport the material in one cycle. The condition of trucks might impact the result, for instance, the age of the trucks or tire condition. However, no documentation can be shared about it for this research. Therefore, this research assumes trucks are in good condition and not have a difference.

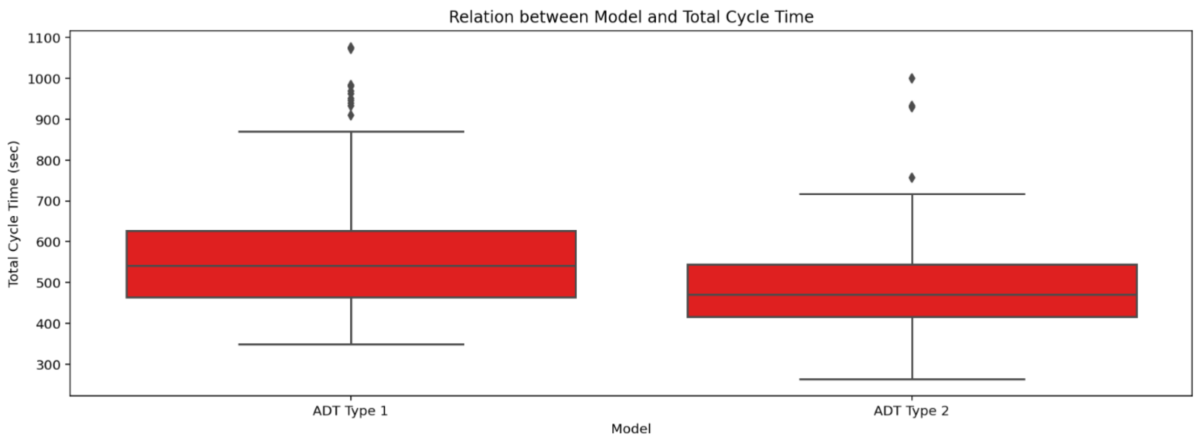


Figure 32. Box Plot of Model from Automated data

Machinery Condition: Weight and Volume

Figure 33 and Figure 34 show the data point of weight and volume of material transported and recorded in Automated data. The range of weight is between 2.6 and 63.2 tonnes, where the maximum weight is 45.2 Tonnes. It shows that many data point exceeds the maximum weight capacity. The range of volume is between 1.1 and 27.7 m³, where the maximum volume capacity is 25. There is only one data point that exceeds the maximum capacity. The value is related to the operation practice in filling the truck bucket with the material. The operator fills the bucket until it is full, not measure the weight.

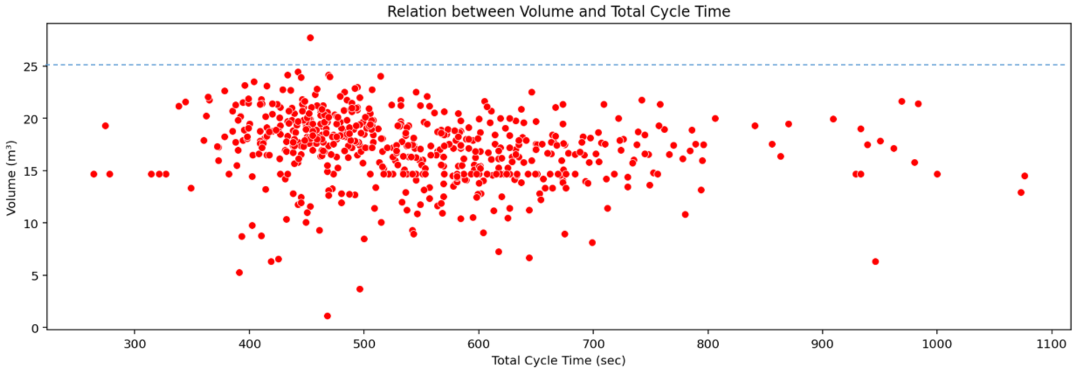


Figure 33. Scatter Plot of Weight in Automated data

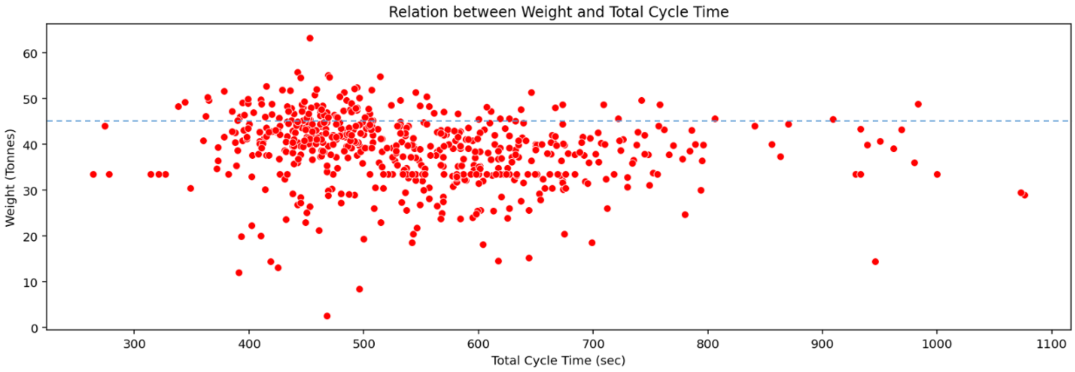


Figure 34. Scatter Plot of Volume in Automated data

3.2.2.2. Exploratory Data Analysis (EDA)

The result from the IDA is analyzed based on the relation between variables and the regression line in the automated data using a pairs plot shown in Figure 35. The five rows from the right are the relation variables with load time, haul time, return time, unload time, and truck cycle time. The data distribution is less spread on the load and unload time rows than a haul, return, and truck cycle time. Most of the data is distributed between zero and 400 seconds for load time and between zero to 60 seconds for unload times. The confidence interval is shown in the area where less data is distributed above 400 seconds for load time and 60 seconds for unload time. It indicates high uncertainty of the result from the regression line of the load time and the unload time.

Moreover, the confidence interval and regression line between variable distance, relative humidity, weight, and volume with the load time is estimated to go downright. It indicates that the value of each variable which small, will cause a longer time to load the material. The confidence interval and regression line between the variable model and the load time is estimated to go upright, indicating that model one will cause a longer time to load the material.

The confidence interval between variables and variable unload time is wide above and under the regression line, indicating that the regression line result can be overestimated or underestimated with high uncertainty. In addition, column weight and volume have a similar pattern but different values. The pairs plot will be analyzed further using the correlation matrix.

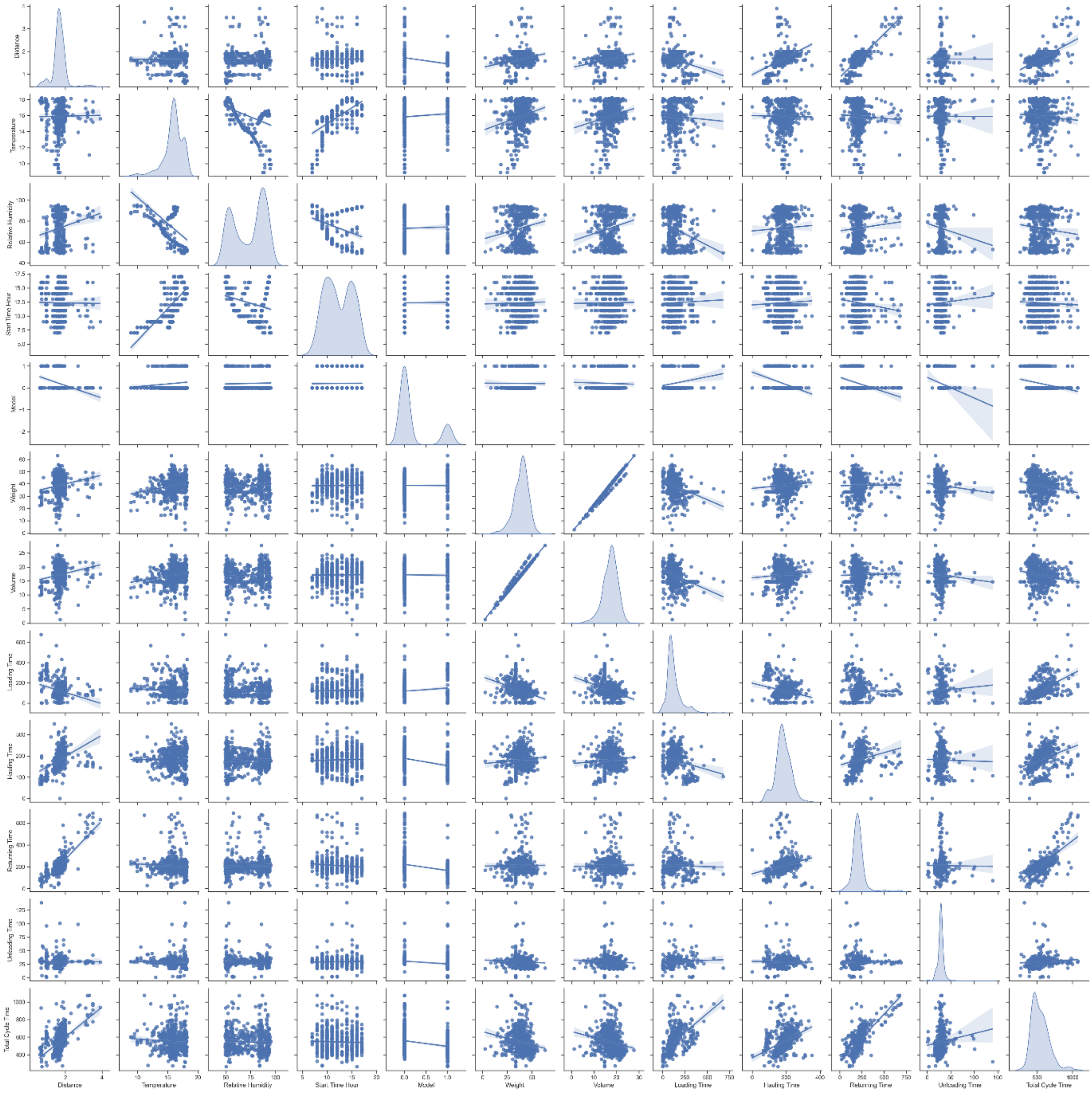


Figure 35. Pairs plot of the automated data

Figure 36 shows the heat map of the correlation matrix, where the calculated value is presented in each box. The matrix shows that weight and volume have a high positive correlation with 0.99, which means the weight and volume value patterns are similar. Hence, ML will not learn significantly from one variable if both variables are included as the input variables. Furthermore, it causes inefficiency of the ML since it takes more time for ML to learn the same pattern. Therefore, it is justified to eliminate one variable and pick one between weight and volume to be the input. Variable volume is selected as one of the feature variables because it proves that the operator is more concerned with the volume of material carried by truck than weight.

Moreover, the matrix shows weight and volume negatively correlated with LT, HT, RT, UT, and TCT. The result indicates a decreasing monotonic trend between variables. However, in theory, it should be a positive correlation which indicates an increasing monotonic trend. Thus, this finding will be assessed from expert perspectives.

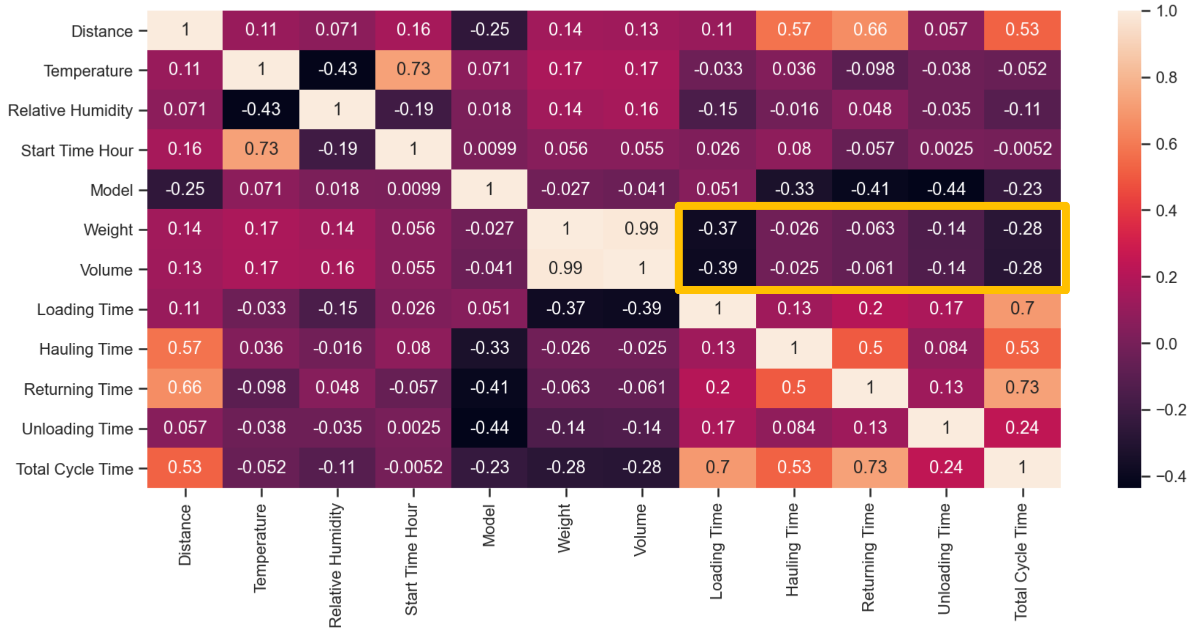


Figure 36. Spearman's correlation matrix of automated data

Figure 37 shows the principal component analysis (PCA) from the automated data. It shows that the principal components contribute approximately 56% of the total variation in the data. PC 1 explains 33.58%, and PC 2 explains 22.89% of the variation in the automated data. PC 1 is influenced by variable temperature, start time hour, and relative humidity. And PC 2 is mainly influenced by variable volume, distance, and model.

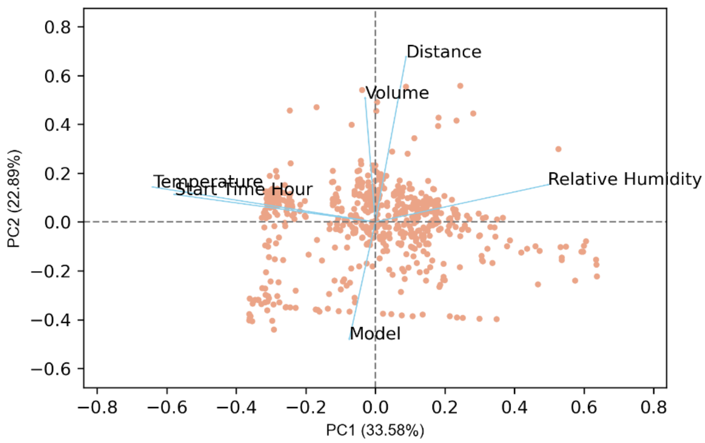


Figure 37. PCA of automated data

3.3. Scenario

Based on the data analysis process, each can develop predictive models with different feature combinations derived from PCA analysis. Variables that influence PC two will be used as the feature combination two, and variables that iinfluencePC one will be used as the feature combination three. The feature combination one will consist of the complete variables. The division of feature combination is applied in the automated data and manual data.

Manual Data

The manual data can be used to develop a predictive model of TCT without RT. The input of the predictive model can use three different feature combination, as follows.

- Combination one: distance, relative humidity, temperature, and start time hour
- Combination two: distance and temperature
- Combination three: relative humidity, temperature, and start time hour

Automated Data

The availability of individual activity duration data in automated data can predict TCT with three different scenarios. The scenarios aim to find the accurate prediction of different predictive models.

- Scenario 1: LT + HT + UT + RT

This scenario requires predictive models of individual activity duration, such as LT, HT, UT, and RT. The prediction result from each will be accumulated and evaluated.

- Scenario 2: TTT + LT + UT

The second scenario requires a new variable which is truck travel time (TTT). Equation 8 is the calculation formula for creating TTT, which is the sum of transporting activity.

Equation 8. TTT of Automated data

$$TCT = HT + RT$$

The predictive model of TTT can help estimate the big part of TCT without considering other equipment, for instance, excavator. The prediction result from TTT will be accumulated with the prediction result from LT and UT.

- Scenario 3: TCT

The second scenario requires a new variable which is TCT. Equation 9 is the formula for creating variable TCT, which is the sum of individual activity duration. This scenario might help to predict TCT more accurately and faster.

Equation 9. TCT of Automated data

$$TCT = LT + HT + UT + RT$$

Therefore, the ML approach will develop LT, HT, UT, RT, TTT, and TCT models. Each model will be developed using three different feature combinations, as follows.

- Combination one: distance, relative humidity, temperature, start time hour, model, and volume
- Combination two: distance, model, and volume
- Combination three: relative humidity, temperature, and start time hour.

4. Predictive Modelling

The previous chapter explained the data preparation phase, where manual and automated data are collected and analyzed into a cleaned data. The result led to different scenarios to achieve output and feature combinations for developing the predictive model. This chapter aims to develop predictive models using a machine learning approach. Hence, this chapter will explain the process of each predictive model from manual and automated data.

4.1. Input and Output

Figure 38 presents the scheme to develop predive modelling. Manual data and automated data are divided into features or input and the target or the output. The output from manual data is TCT exclude RT. The output will be developed using feature combinations. The output from automated data is TCT, TTT, LT, HT, UT, and RT. Individual output is required to combine the prediction result for obtaining TCT. The output will be developed using three feature combinations from automated data.

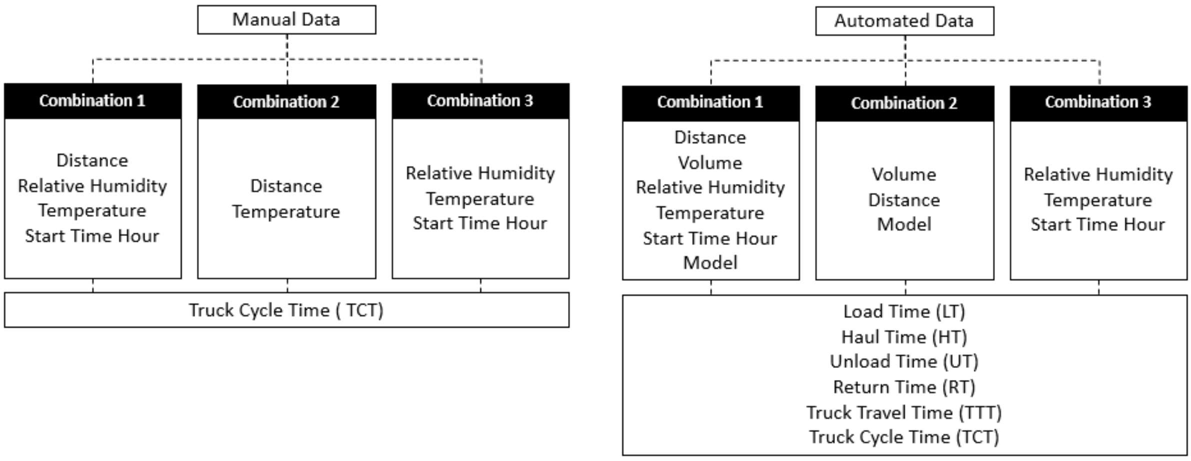


Figure 38. Scheme to develop the predictive model

4.2 Train, Validation, and Test Dataset

The machine learning approach requires the dataset to have a training dataset and a test dataset. The training dataset is used for training the model to find the prediction pattern. The prediction modelling performance is measured using the test dataset. Train data set and test dataset consist of 80% and 20% of the total data point. The data split randomly for avoiding bias model or result.

The training data set is divided into the training dataset and the validation dataset. The validation dataset or development data set is used for developing the model by tuning the parameters in the model. Tuning the parameters for a model is through a highly iterative process that starts with an idea, finds the code, and does an experiment until finding a better result [21]. Therefore, the training dataset and validation dataset need to be split properly for obtaining a better result.

This research aims to achieve a good performance model that gives an accurate prediction with future input data. Figure 39 shows three possible predictive model performances: under-fitting, good fit, and overfitting. Underfitting refers to a model that not able to generalize data or find a fit pattern. Overfitting is a model that learns the data in detail and poorly predicts the new data because it remembers the train data. And good fit/robust, which is the desired outcome, refers to a model that learns data generally and predicts new data.

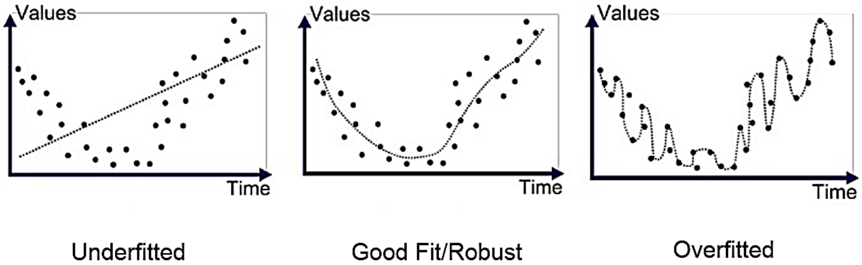


Figure 39. Illustration of Underfitted, Good Fit, and Overfitted [22]

This research will use the k-fold cross-validation in splitting and iterating the training data, so the predictive model is ensured not to have an overfitted performance. The explanation of k-fold cross-validation will be explained in the following section.

4.3. Regression Technique

Based on data analysis, the relationship between variables and target can be linear and non-linear. Hence, this research will use different types of regression algorithms to develop the predictive model of TCT, which is a regression problem. This research will develop predictive models by starting with a simple regression algorithm from traditional ML, such as multiple linear regression (MLR) and support vector regression (SVR), and continuing with a more complex algorithm from DL, such as Artificial Neural Network (ANN).

Each regression technique has a different approach to develop a regression model, such as MLR uses Ordinary Least Squares, SVR uses hyperplanes, and ANN uses multilayers. Those regression techniques will develop models from different feature combination as the input. The results from different feature combinations and regression techniques might give insights and understanding about predictive models.

Algorithms will develop the regression model in python with different library package. Hence, a manual calculation is not required to develop the regression model. The following is a brief explanation of each regression technique that will be used.

4.3.1. Multiple Linear Regression

Multiple Regression or Multiple Linear Regression (MLR) is a statistical technique that uses explanatory variables to predict the outcome. The basic form of MLR is simple linear regression which is illustrated in Figure 40. Linear regression functions well in predicting linear data set and only fits one dependent and one independent variable. Equation 11 is the formula for simple linear regression, where y is the dependent variable, x is the independent variable, $\hat{\beta}$ is coefficient and ε is the intercept value that dictating the equation.

Equation 10. Simple Linear Regression

$$y = \hat{\beta}x + \varepsilon$$

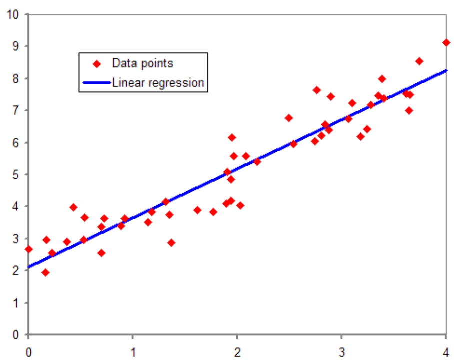


Figure 40. Example of Simple Linear Regression.

However, it will spend times and effort for making one by one prediction model for different variables. Thus, MLR is useful as an efficient way for predicting one dependent variable and multiple independent variables. The general form of the equation for multiple regression, as follows. Where $\hat{\beta}_1$ refers to the first independent variables until n number of input variables. In linear regression,

Equation 11. Multiple Linear Regression

$$y_i = \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_n x_{in} + \varepsilon_i$$

A regression model has good performance if it has a minimum value of the sum of squared residual. Least squares is an approach in regression analysis that minimize the sum of the squares of the residuals. Two categories of least squares are linear or ordinary least squares and non-linear squares. Equation 12 shows Ordinary Least Squares (OLS) where y_i is the target, $w_i$is the coefficient, and $x_i$is the input or feature.

Equation 12. Ordinary Least Squares (OLS)

$$MIN \sum_{i=1}^{n} (y_i - w_i x_i)^2$$

4.3.2 Support Vector Machine

Support Vector Regression (SVR) is a regression analysis using the Support Vector Machine (SVM) method. SVM develops predictive models by constructing hyperplanes for solving classification or regression problem. Figure 41 shows how SVR works by considering points within the boundary line (grey line) and minimizing error. Equation 13 is used for calculating the boundaries where y_i is the target, $w_i$is the coefficient, and $x_i$is the predictor or feature, and ε is is a margin of tolerance . Equation 14 is used for minimizing the error.

Equation 13. Boundary for SVR

$$|y_i - w_i x_i| \leq \varepsilon$$

Equation 14. Minimize error

$$MIN \ \frac{1}{2} ||w||^2$$

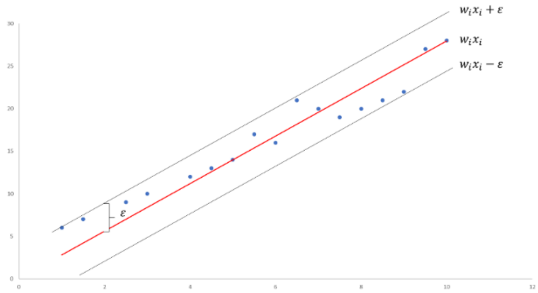


Figure 41. Example of Support Vector Regression

4.3.3.Artificial Neural Network

Artificial Neural Network (ANN) which is part of DL, imitates how the biological neural networks process the information by processing the input data into layers. The system will find the appropriate answer [23]. ANN has a collection of neuron, which is called artificial neurons. Each neuron is connected with the edges. Like the brain activity where a signal is transmitted to neurons, the artificial neurons also get a signal transmitted through the connection. ANN has two main elements, which are knowledge and the interneuron connection strength [23]. Knowledge obtains through the learning process of the machine. Interneuron connection strength is needed to store the knowledge. The machine learns from its environment, which consists of three different types of layers, as follows.

- The input layer is a layer that receives the data as the input.
- Hidden layers are layers to optimize the weight to improve the prediction result.
- The output layer is a layer that gives the prediction result.

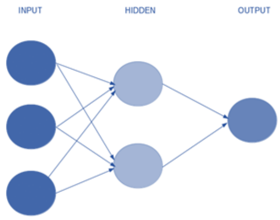


Figure 42. Artificial Neural Network Structure

Figure 42 illustrates the connection between the input, hidden, and output layer in the ANN structure. The number of hidden layer and number of neurons can be added more than one. The structure aims to process the knowledge from data and create a predictive model. Figure 43 shows the learning process of ANN, which contains two main parts, such as front propagation (FP) and backpropagation (BP). FP refers to input data processing that passes through neuron layers in a neural network from the input to the output layers. BP refers to propagating the error back into the network and updating each weight and bias.

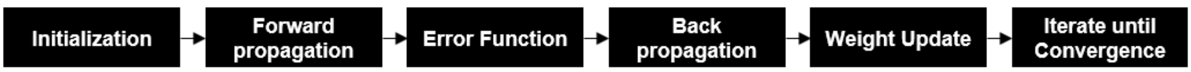


Figure 43. The process of ANN [24]

Front Propagation

Initial data is propagated through network architecture structured by its depth, width, and activation function in each layer [25]. Depth is defined by the number of hidden layer in the network. Width is defined by the number of neurons of nodes that is applied in each layer. An activation function calculates the sum of neuron weight in the input layer, adds the bias, and transforms it into the output layer. Figure 44 is a sample of how an activation function works in a neuron.

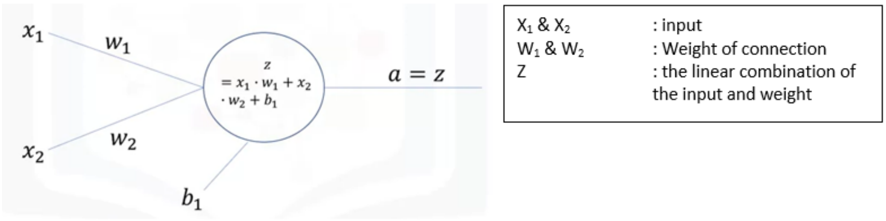


Figure 44. Sample of Front Propagation in a neuron

Activation Function

An activation function is chosen for the hidden layer and output layer because they serve different objectives. Many activation functions can be applied in ANN, such as Linear, Sigmoid, Rectified Linear (ReLU). ReLu has a function which returns 0 for a negative value. Otherwise, the value is returned. ReLU is the most used activation function for hidden layers because it can overcome the vanishing gradient problem, allowing the model to learn faster and better performance [21]. Therefore, this research uses ReLu for the hidden layers.

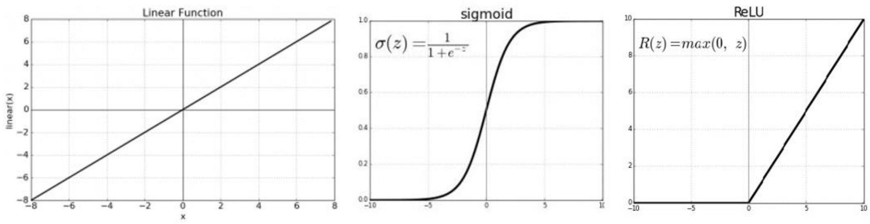


Figure 45. Activation Function: Linear, Sigmoid, and ReLu

Figure 45 shows different activation functions for the output layers, such as ReLU, Sigmoid, and Linear Sigmoid activation function, which takes a real value as input and output values in the range between 0 for negative value and 1 for a positive value. A linear activation function is also known as no activation function because it does not change the weighted sum and return the value directly.

Hyperparameter

Besides deciding activation functions, batch size and epoch must be set before the learning process begins, called hyperparameter. Batch size is the number of training samples that are used to be run in one iteration. A larger batch size will increase the speed of the learning rate, but it can decrease the accuracy. Epoch refers to the number of a cycle that through training dataset.

Backpropagation

After receiving the result from the forward propagation process, the backpropagation algorithm calculates the gradient of the lost function to each weight. Figure 46 shows the illustration of the backpropagation process with respect to the neural network weight. The loss function and optimizer is required to be selected for ANN structure to conduct backpropagation.

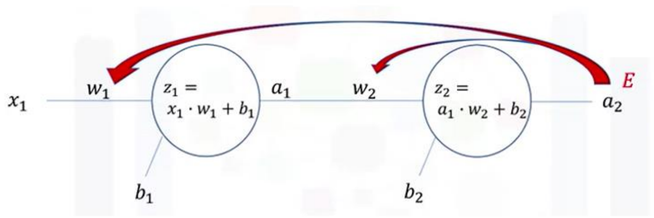


Figure 46. Sample of Backpropagation

Loss Function

The loss function is a mathematical algorithm that helps measure the model's performance toward the desired result. The typical loss function used for the regression problem is Mean Squared Error (MSE). Equation 15 is the function of MSE where $\hat{y}$ is predicted value of y and $\bar{y}$ is the mean value of y. This research will use MSE as the loss function because MSE is a sensitive calculation for a big range of loss.

Equation 15. Mean Squared Error

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

Optimizer

The optimizer is a mathematical algorithm that helps the loss function reach its peak performance without delay and provide the most accurate result. Adaptive Moment Estimation (Adam) is one of the optimizer types commonly used in a neural network. Adam optimization is a stochastic gradient descent method based on adaptive estimation for each parameter [26].

4.4. K-Fold Cross-Validation

The training dataset is split using the k-fold cross-validation method, a common method in machine learning to overcome the overfitting problem [27]. Figure 47 illustrates the process of k-fold cross-validation in a dataset. In the beginning, the training dataset is shuffled randomly and split into a certain number of k. Each fold will select a group of data to be a validation set and use the remaining data as the training set. The iteration process decides the number of k until a better outcome for the model is found. Finally, the outcome is evaluated with performance metric in this research.

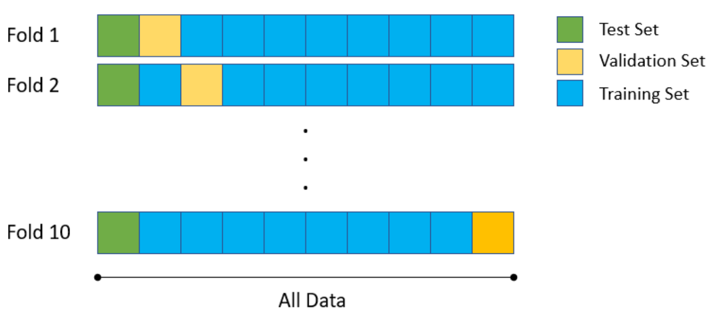


Figure 47. Illustration of K-fold 10

4.5. Performance Metrics

The model's performance is evaluated using common three statistical methods in ML, such as RMSE, MAE, and R^2. RMSE or Root Mean Squared Error is the error rate of the square root of the difference between the original and predicted value extracted by squared the average difference over the data set [28]. MAE or Mean Absolute Error is the error value calculated from the absolute difference between the original and predicted values over the dataset. R^2, or the coefficient of determination, is a measurement to explain the variability of a variable to another variable. The following is the function of the methods where $\hat{y}$ is predicted value of y and $\bar{y}$ is the mean value of y

Equation 16. Root Mean Squared Error (RMSE)

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

Equation 17. Mean Absolute Error (MAE)

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

Equation 18. Coefficient of Determination (R^2)

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

The measurement has different purposes in evaluating predictive models. RMSE calculation is sensitive to poor prediction, and MAE treats all the prediction error proportionally. It also the reason why MAE is more robust to cope with outliers than RMSE. The value of RMSE and MAE that closer to zero indicates better model performance. R^2 calculation aims to know how many data points that fall within the model. R^2 value that closes to one indicates better model performance. The value can refer to a percentage of the goodness of fit. In addition, R^2 can be negative, which indicates the model has poor goodness of fit toward the data distribution or the intercept (ε) of the MLR from Equation 11 haven't set. This research will set the intercept to obtain the best model.

The modelling might have a trade-off, for instance, low RMSE and high R^2 but high MAE. Then, the predictive model that contained less error and high goodness of fit value will be selected. Therefore, each value is considered in the modelling process, where the detailed process is shown in Appendix 2.

4.6. Modelling

Figure 48 shows the scheme of using the dataset in the modelling process. However, this research has limitation regarding the size of the dataset, which is not big enough, time, and the capacity of the computer for developing models. Hence, this research does not use a high parameter tuning where models are not developed with a big number for each parameter. Instead, each predictive models will be developed between two and ten folds in the k-fold cross-validation process.

Predictive models with ANN method are developed by tuning the batch size 10, 50, 100, 150, epochs within 10, 50,100,500, and 1000, neurons within one, four, 12, 36 for manual data and 1, 6, 12, 24, and 36 for automated data, and hidden layers between one to five. Then, the selected predictive models will be tested with the test dataset.

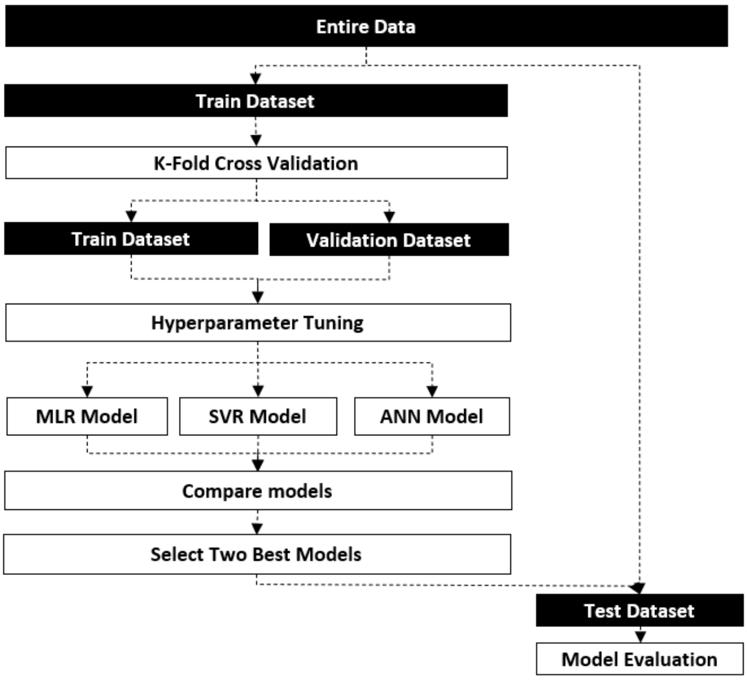


Figure 48. Scheme of using the data set

4.6.1. Manual data

The data points in manual data are divided into two parts, where 344 data points for the training dataset and 84 data points for the test dataset. The following is the outcome of predictive models with different methods and different feature combination. The detailed development process for each predictive model is presented in Appendix 2.

Multi Linear Regression

Table 7 shows the coefficient and intercept value of each TCT predictive model that uses the Multi Linear Regression (MLR) method and different feature combinations from manual data. Each model is developed with different folds. The number of folds for combination one is seven-folds, combination two is five-folds, and combination three is two-folds. The coefficient and intercept value affect the model performance.

Table 7. Multi Linear Regression result from manual data

Coefficient	Combination one	Combination two	Combination three
Intercept	0.068	0.086	0.072
Start Time Hour	0.00004	-	-0.024
Distance	0.0308	0.041	-
Temperature	0.0499	0.037	0.077
Relative Humidity	0.0235	-	0.027

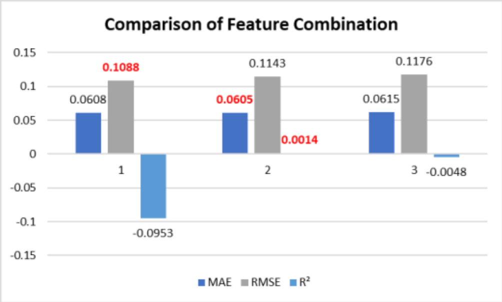


Figure 49. Comparison of TCT models from manual data using MLR.

Figure 49 presents that MLR obtains the lowest RMSE value with 0.1088 using combination one, the lowest MAE value with 0.0605 and the highest R^2 value with 0.004 using combination two. It indicates that MLR with combination two develops a TCT model which obtains the lowest average error and can capture most data points compare with other models. However, the model is considered to have low accuracy because the goodness of fit is close to zero.

Support Vector Regression

Figure 50 shows the comparison between TCT predictive models that use the Support Vector Machine (SVR) method and different feature combinations from manual data. The number of folds for combination one is eight-folds, combination two is seven-folds, and combination three is two-folds.

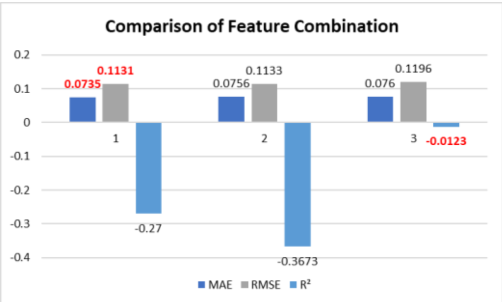


Figure 50. Comparison TCT models from manual data using SVR

The result presents that SVR obtains the lowest MAE value with 0.0735 and the lowest RMSE value with 0.1131 using combination one and the highest R^2 value with -0.0123 using combination three. It indicates that SVR with feature combination one develops a TCT model, which obtains the lowest average error and has few data points far from the model. However, the model cannot capture most data points.

Artificial Neural Network

Figure 51 shows the comparison between TCT predictive models that use Artificial Neural Network (ANN) method and different feature combinations from manual data. First, combination one is developed by applying ten batch size, 100 epochs, 36 neurons in four hidden layers, and three-folds. Next, combination two is developed by applying the ten batch size, 500 epochs, 36 neurons in two hidden layers, and four-folds. Finally, combination three is developed by applying the ten batch bize, 100 epochs, 36 neurons in five hidden layers, and three-folds.

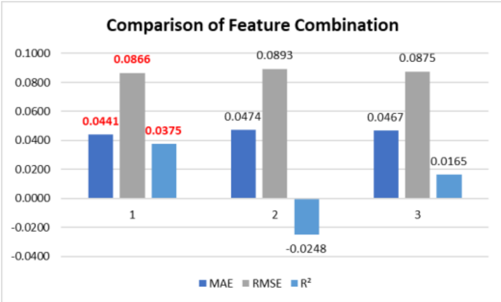


Figure 51. Comparison of TCT models from manual data using ANN

The result presents that ANN obtains the lowest MAE value with 0.0441, the lowest RMSE value with 0.0866, and the highest R^2 value with 0.0375 using combination one. It indicates that ANN with combination one develops a TCT model, which has the lowest average error, close to most data points, and can capture most data points.

Overview

Figure 52 shows the comparison between methods in each combination with manual data. Combination one obtains the best predictive model with MLR, with the lowest MAE and RMSE values, although ANN has the highest R^2 value. It indicates that combination one is suitable to develop a regression model using the MLR method. It can obtain prediction with a small average error and is not far from the data points but cannot capture most data points. With combination two, MLR has the lowest MAE value and the highest R^2 value, but ANN has the lowest RMSE value. It indicates that combination two is suitable to develop a regression model using MLR. It can capture most data points and low average error, but it is far from most data points. With combination three, MLR has the lowest MAE, RMSE value and the highest R^2 value. It indicates that combination three is suitable for developing a regression model using the MLR method. The model can capture most data points, obtain a low average error, and close to most data points.

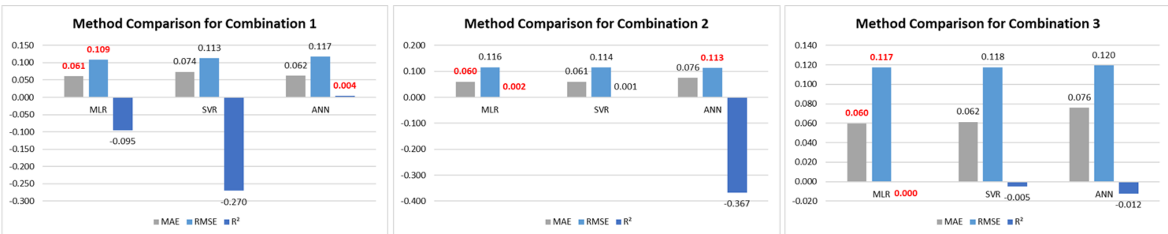


Figure 52. TCT models from manual data.

Overall, predictive models with manual data cannot capture most data points. The plausible reason is due to the bad quality of data. The data collected manually is influenced by many factors that might affect the result and the data pattern. Therefore, the methods are difficult to develop a reliable predictive model. This result also is influenced by the data quantity where the training dataset consists of 344 data points. Thus, the predictive models from manual data will not be analyzed further because the models are insufficient to be used as TCT predictive models in earthworks.

4.6.2. Automated data

This section's main objective is to develop a predictive model of TCT, individual activity duration, and Truck Travel Time (TTT). The data points in automated data are divided into two parts, where 471 data points for the training dataset and 118 data points for the test dataset. The following is the outcome of predictive models with different methods and different feature combination. The detailed development process for each predictive model and the mathematic equation from MLR are presented in the Appendix.

4.6.2.1. Load Time

The result from load time (LT) predictive models with different methods and feature combinations will be explained.

Multi Linear Regression

Figure 53 shows the comparison between LT predictive models that use Multi Linear Regression (MLR) method and different feature combinations from automated data. The number of folds for combination one and two is three-folds, and combination three is ten-folds.

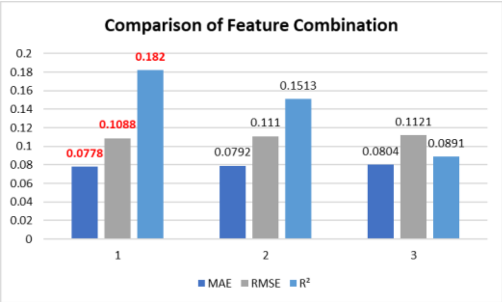


Figure 53. Comparison of LT models from automated data using MLR

The result shows MLR obtains the lowest MAE value with 0.0778, the lowest RMSE value with 0.1088, and the highest R^2 value with 0.182 using combination one. It indicates that MLR obtains an LT model with the lowest average error, close to most data points, and can capture most data points with combination one.

Support Vector Regression

Figure 54 shows the comparison between LT predictive models that use the Support Vector Machine (SVR) method and different feature combinations from automated data. The number of folds for combination one is two-folds, combination two is nine-folds, and combination three is nine-folds.

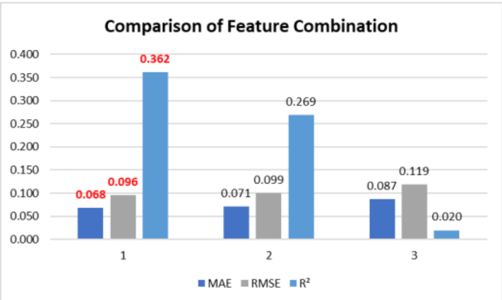


Figure 54. Comparison of LT models from automated data using SVR

The result presents that SVR obtains the lowest MAE value with 0.068, the lowest RMSE value with 0.096, and the highest R^2 value with 0.362 using combination one. It indicates that SVR and combination one develop an LT model which obtains the lowest average error, close to most data points, and can capture most data points.

Artificial Neural Network

Figure 55 shows the comparison between LT predictive models that use Artificial Neural Network (ANN) method and different feature combinations from automated data. First, combination one is developed by applying ten batch size, 500 epochs, 36 neurons in one hidden layer, and three-folds. Next, combination two is developed by applying the ten batch size, 500 epochs, 12 neurons in one hidden layer, and three-folds. Finally, combination three is developed by applying the ten batch bize, 500 epochs, 12 neurons in one hidden layer, and ten-folds.

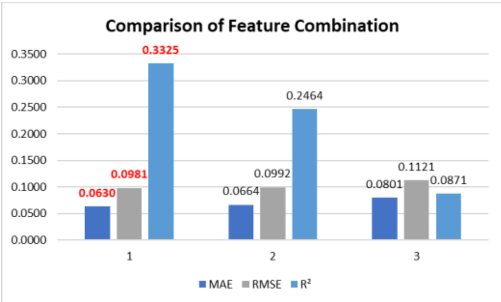


Figure 55. Comparison of LT models from automated data using ANN

The result presents that ANN obtains the lowest MAE value with 0.0630, the lowest RMSE value with 0.0981, and the highest value of R^2 with 0.3325 using combination one. It indicates that ANN and combination one develop an LT model which obtains the lowest average error, close to most data points, and can capture most data points.

Overview

Figure 56 shows the comparison between methods in each combination with automated data. Combination one and two obtain the best predictive model with SVR, with the lowest RMSE and the highest R^2 value, although ANN has the lowest MAE value. Hence, it indicates that combination one is suitable to develop a regression model using the SVR method. Each predictive model is close to the data points and can capture most data points, although the average error is not the lowest.

With combination three, the lowest RMSE value is obtained using MLR and ANN method. MLR also has the highest R^2 value, and ANN has the lowest MAE value. It indicates that the input from combination three develops similar predictive models using MLR and ANN. Both regression models are close to most data points. MLR model can capture most data points, and the ANN model has the lowest average error.

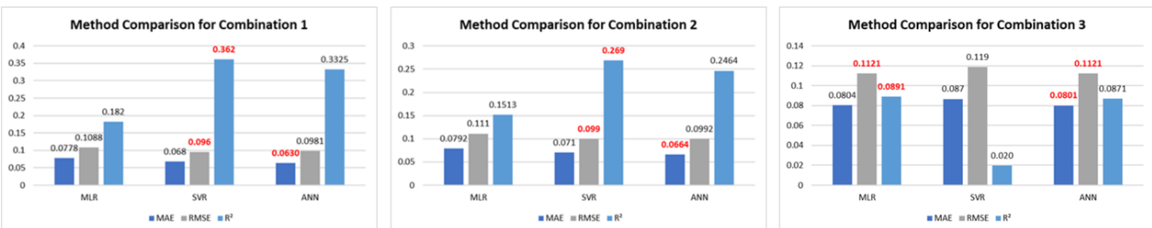


Figure 56. Comparison of LT models from automated data

Overall, the best LT model is developed by SVR with feature combination one from automated data. It has a similar performance with the predictive model from ANN with feature combination one.

4.6.2.2. Haul Time

The results from haul time (HT) predictive models with different methods and feature combinations will be explained.

Multi Linear Regression

Figure 57 shows the comparison between HT predictive models that use Multi Linear Regression (MLR) method and different feature combinations from automated data. The number of folds for combination one and two is three-folds, and combination three is ten-folds.

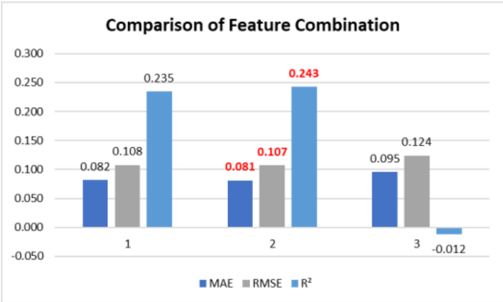


Figure 57. Comparison of HT models from automated data using SVR

The result shows MLR obtains the lowest MAE value with 0.081, the lowest RMSE value with 0.107, and the highest R^2 value with 0.243 using combination two. It indicates that MLR and combination two develop an HT model which obtains the lowest average error, close to most data points, and can capture most data points.

Support Vector Regression

Figure 58 shows the comparison between HT predictive models that use the Support Vector Machine (SVR) method and different feature combinations from automated data. The number of folds for combination one, two, and three is nine-fold.

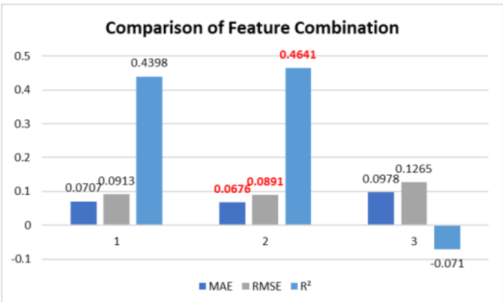


Figure 58. Comparison of LT models from automated data using SVR

The result presents that SVR obtains the lowest MAE value with 0.068, the lowest RMSE value with 0.089, and the highest R^2 value with 0.464 using combination two. It indicates that SVR and combination one develops an HT model that obtains the lowest average error, close to most data points, and can capture most data points.

Artificial Neural Network

Figure 59 shows the comparison between HT predictive models that use Artificial Neural Network (ANN) method and different feature combinations from automated data. First, combination one is developed by applying ten batch size, 500 epochs, 24 neurons in four hidden layer, and seven-folds. Next, combination two is developed by applying the ten batch size, 1000 epochs, 24 neurons in three hidden layers, and ten-folds. Finally, combination three is developed by applying the ten batch bize, 100 epochs, 6 neurons in one hidden layer, and nine-folds.

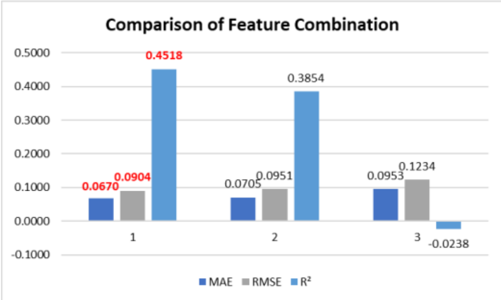


Figure 59. Comparison of HT models from automated data using ANN

The result presents that ANN obtains the lowest MAE value with 0.0670, the lowest RMSE value with 0.0904, and the highest value of R^2 with 0.4518 using combination one. The result indicates that ANN and combination one develop an HT model which obtains the lowest average error, close to most data points, and can capture most data points.

Overview

Figure 60 shows the comparison of HT models from different methods in each combination with automated data. Combination one obtains the best predictive model with ANN, with the lowest MAE and RMSE values and the highest R^2 value. It indicates that combination one is suitable to develop a regression model using the ANN method. It can predict with a small average error, not far from the data points, and capture most data points. With combination two, SVR has the lowest MAE, RMSE value and the highest R^2 value. It indicates that combination two is suitable to develop a regression model using SVR. It can capture most data points, has a low average error, and close from most data points. With combination three, MLR has the lowest MAE value and the highest R^2 value, but ANN has the lowest RMSE value. It indicates that combination three is suitable for developing a regression model using the MLR method. The model can capture most data points, obtain a low average error, but far from most data points.

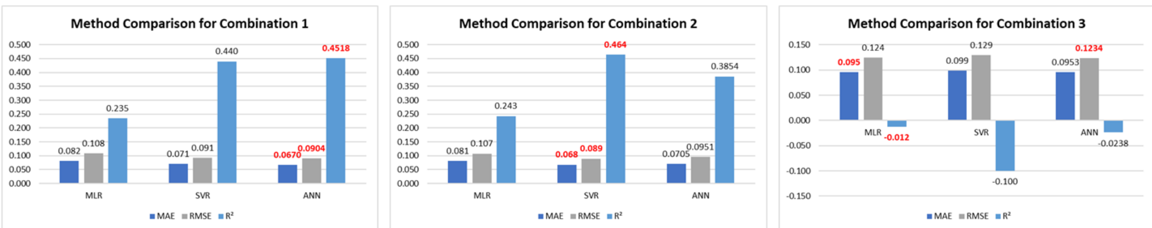


Figure 60. Comparison of HT models from automated data

Overall, the best HT model is developed by SVR with feature combination two from automated data. It has a similar performance with the predictive model from ANN with feature combination one.

4.6.2.3. Unload Time

The following results from predictive models of unload time (UT) with different methods and feature combinations.

Multi Linear Regression

Figure 61 shows the comparison between UT predictive models that use Multi Linear Regression (MLR) method and different feature combinations from automated data. The number of folds for combination one and two is seven-folds, and combination three is ten-folds.

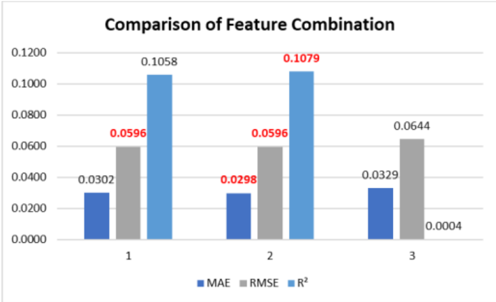


Figure 61. Comparison of UT models from automated data using MLR

The result shows MLR obtains the lowest MAE value with 0.0298, the lowest RMSE value with 0.0596, and the highest R^2 value with 0.1079 using combination two. It indicates that MLR with combination two develops a UT model which obtains the lowest average error, close to most data points, and can capture most data points.

Support Vector Regression

Figure 62 shows the comparison between UT predictive models that uses the Support Vector Machine (SVR) method and different feature combinations from automated data. The number of folds for combination one is three-folds, combination two is two-folds, and combination three is eight-folds.

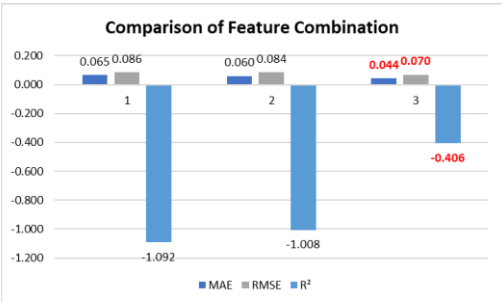


Figure 62. Comparison of UT models from automated data using SVR

The result presents that SVR obtains the lowest MAE value with 0.044, the lowest RMSE value with 0.070, and the highest R^2 value with -0.408 using combination three. It indicates that SVR and combination three develop a UT model which obtains the lowest average error, close to most data points, and can capture most data points.

Artificial Neural Network

Figure 63 shows the comparison between UT predictive models that use Artificial Neural Network (ANN) method and different feature combinations from automated data. First, combination one is developed by applying ten batch size, 100 epochs,6 neurons in one hidden layer, and seven-folds. Next, combination two is developed by applying the ten batch size, 100 epochs, 6 neurons in one hidden layer, and six-folds. Finally, combination three is developed by applying the 50 batch bize, 1000 epochs, three neurons in one hidden layer, and three folds. The result presents that ANN obtains the lowest MAE value with 0.0291, the highest value of

R^2 with 0.1340 using combination two, and the lowest RMSE value with 0.0589 using Combination one and 2.

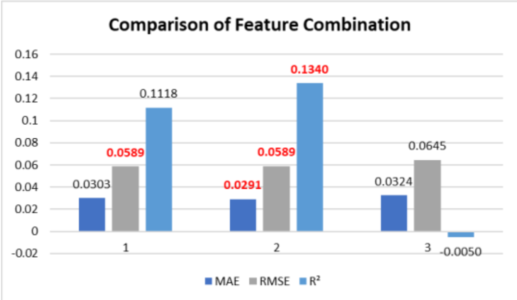


Figure 63. Comparison of UT models from automated data using ANN

The result presents that ANN obtains the lowest MAE value with 0.0291, the highest value of R^2 with 0.1340 using combination two, and the lowest RMSE value with 0.0589 using Combination one and two. It indicates that ANN with combination two develops a UT model that obtains the lowest average error, is close to most data points, and can capture most data points.

Overview

Figure 64 shows the comparison of UT models from different methods in each combination with automated data. Combination one obtains the best predictive model with ANN, with the lowest RMSE and the highest R^2 value, although MLR has the lowest MAE values. It indicates that combination one is suitable to develop a regression model using the ANN method. It is not far from the data points and can capture most data points. With combination two, ANN has the lowest MAE, RMSE value and the highest R^2 value. It indicates that combination two is suitable to develop a regression model using ANN. It can capture most data points, has a low average error, and close from most data points. With combination three, MLR has the lowest RMSE value and the highest R^2 value, but ANN has the lowest MAE value. It indicates that combination three is suitable for developing a regression model using the MLR method. The model can capture most data points, obtain a low average error, but far from most data points.

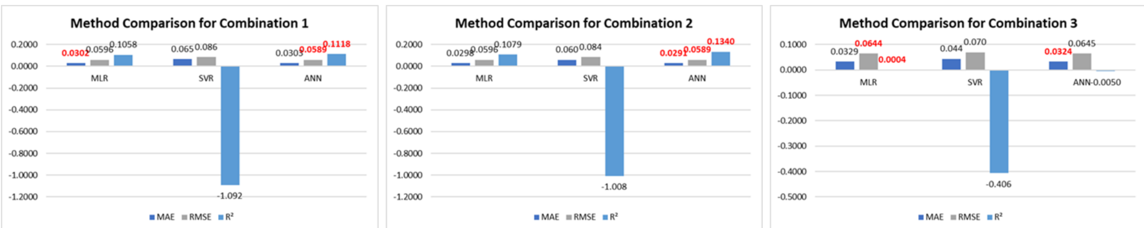


Figure 64. Comparison of UT models from automated data

Overall, the best UT model is developed by ANN with feature combination two from automated data. It has a similar performance with the predictive model from ANN with feature combination one.

4.6.2.4. Return Time

The following results from return time (RT) predictive models with different methods and feature combinations.

Multi Linear Regression

Figure 65 shows the comparison between RT predictive models that use Multi Linear Regression (MLR) method and different feature combinations from automated data. The number of folds for combination one is six-folds, and combination two and 3 are three-folds.

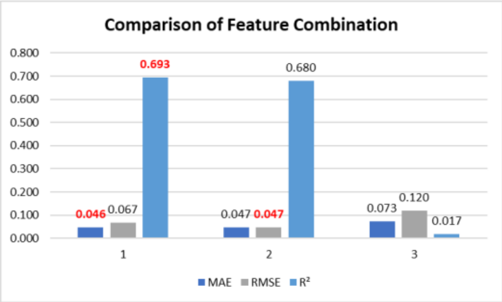


Figure 65. Comparison of RT models from automated data using MLR

The result shows MLR obtains the lowest MAE value with 0.046, the highest R^2 value with 0.693 using combination one, and the lowest RMSE value with 0.047 by combination two. It indicates that MLR with combination two develops an RT model that obtains the lowest average error and can capture most data points. However, the RT model is far from some data points.

Support Vector Regression

Figure 66 shows the comparison between RT predictive models that uses the Support Vector Machine (SVR) method and different feature combinations from automated data. The number of folds for combination one and 2 is six-folds, and combination three is seven-folds.

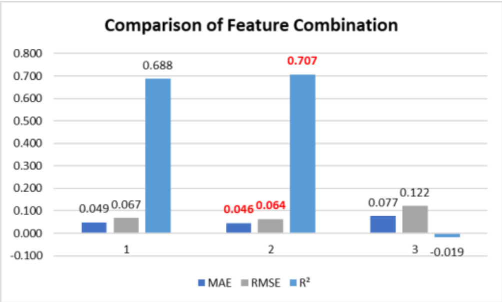


Figure 66. Comparison of RT models from automated data using SVR

The result presents that SVR obtains the lowest MAE value with 0.046, the lowest RMSE value with 0.064, and the highest R^2 value with 0.707 using combination two. It indicates that SVR develops an RT model which has the lowest average error, close to most data points, and can capture most data points with combination two

Artificial Neural Network

Figure 67 shows the comparison between RT predictive models that use Artificial Neural Network (ANN) method and different feature combinations from automated data. Combination one is developed by applying ten batch size, 100 epochs,6 neurons in one hidden layer, and six-folds. Combination two is developed by applying the ten batch size, 100 epochs, 12 neurons in three hidden layers, and six-folds. Combination three is developed by applying the ten batch bize, 100 epochs, 12 neurons in one hidden layer, and two folds. The result presents that ANN obtains the lowest MAE value with 0.0303 and the lowest RMSE value with

0.0589 using combination1, and the highest R^2 value with 0.7203 using feature combination two

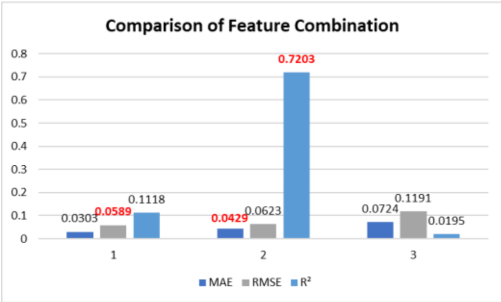


Figure 67. Comparison of RT models from automated data using ANN

Overview

Figure 68 shows the comparison of RT models from different methods in each combination with automated data. Combination one obtains the best predictive model with ANN, with the lowest MAE and RMSE values, although MLR has the highest R^2 value. It indicates that combination one is suitable to develop a regression model using the ANN method. It is not far from the data points and has the lowest average error. With combination two, ANN has the lowest MAE and the highest R^2 value, although MLR has the lowest MSE value. It indicates that combination two is suitable to develop a regression model using ANN. It can capture most data points and has the lowest average error. With combination three, ANN has the lowest MAE, RMSE value and the highest R^2 value. It indicates that combination three is suitable for developing a regression model using the ANN method. The model can capture most data points, obtain a low average error, and close with most data points.

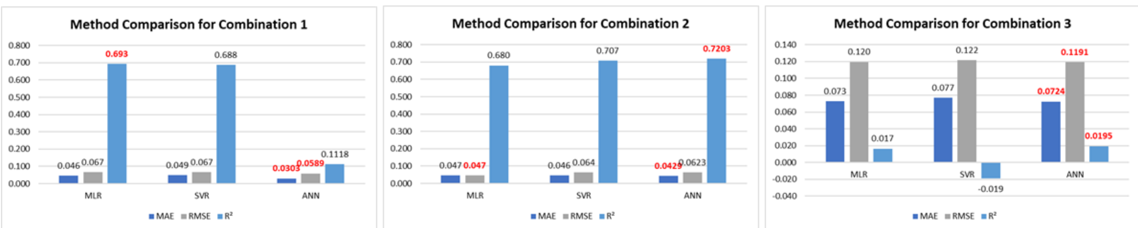


Figure 68. Comparison of RT models from automated data

Overall, the best UT model is developed by ANN with feature combination two from automated data. It has a similar performance with the predictive model from SVR with combination two.

4.6.2.5. Truck Travel Time

The following presents truck travel time (TTT) predictive models with different methods and feature combinations.

Multi Linear Regression

Figure 69 shows the comparison between TTT predictive models that use Multi Linear Regression (MLR) method and different feature combinations from automated data. The number of folds for combination one and two is three-folds, and combination three is two-folds.

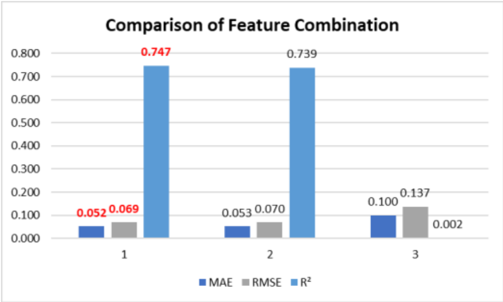


Figure 69. Comparison of TTT models from automated data using MLR

The result shows MLR obtains the lowest MAE value with 0.052, the highest R^2 value with 0.69, and the lowest RMSE value with 0.747 using combination one. It indicates that MLR with combination one develops a TTT model which obtains the lowest average error, close to most data points, and can capture most data points.

Support Vector Regression

Figure 70 shows the comparison between TTT predictive models that use the Support Vector Machine (SVR) method and different feature combinations from automated data. The number of folds for combination one and 3 is seven-folds, and combination two is five-folds.

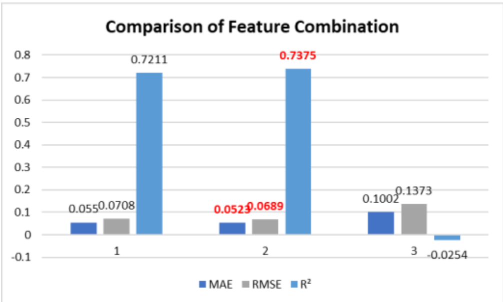


Figure 70. Comparison of TTT models from automated data using SVR

The result presents that SVR obtains the lowest MAE value with 0.0523, the lowest RMSE value with 0.0689, and the highest R^2 value with 0.7375 using combination two. It indicates that SVR develops a TTT model which obtains the lowest average error, close to most data points, and can capture most data points with combination two.

Artificial Neural Network

Figure 71 shows the comparison between TTT predictive models that use Artificial Neural Network (ANN) method and different feature combinations from automated data. First, combination one is developed by applying 100 batch size, 100 epochs, 12 neurons in five hidden layer, and two-folds. Next, combination two is developed by applying the ten batch size, 100 epochs, 12 neurons in five hidden layers, and six-folds. Finally, combination three is developed by applying the ten batch bize, 100 epochs, 36 neurons in two hidden layer, and two folds.

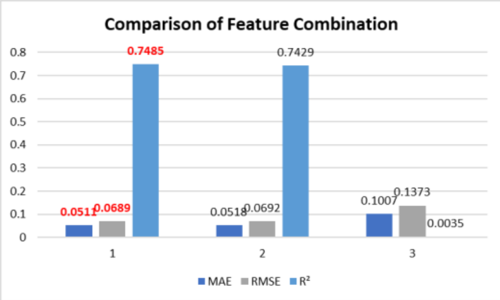


Figure 71. Comparison of TTT models from automated data using ANN

The result presents that ANN obtains the lowest MAE value with 0.0511 and the lowest RMSE value with 0.0689 using combination one, and the highest R^2 value with 0.7485 using combination two. It indicates that ANN develops a TTT model that obtains the lowest average error and can capture most data points, but it is far from some data points, with combination two.

Overview

Figure 68 shows the comparison of TTT models from different methods in each combination with automated data. Combination one obtains the best predictive model with ANN, with the lowest MAE, RMSE, and the highest R^2 value. It indicates that combination one is suitable to develop a regression model using the ANN method. It is not far from the data points, has the lowest average error, and can capture most data points. With combination two, ANN has the lowest MAE and the highest R^2 value, although MLR has the lowest MSE value. It indicates that combination two is suitable to develop a regression model using ANN. It can capture most data points, has the lowest average error. With combination three, ANN has the lowest RMSE value and the highest R^2 value. It indicates that combination three is suitable for developing a regression model using the ANN method. The model can capture most data points and close from most data points.

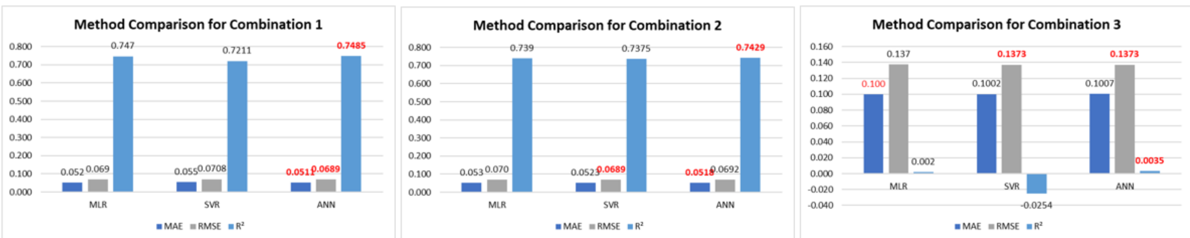


Figure 72. Comparison of TTT models from automated data

Overall, the best TTT model is developed by ANN with feature combination one from automated data. It has a similar performance with the predictive model from MLR with combination one.

4.6.2.6. Truck Cycle Time

The following results from truck cycle time (TCT) predictive models with different methods and feature combinations.

Multi Linear Regression

Figure 73 shows the comparison between TCT predictive models that use Multi Linear Regression (MLR) method and different feature combinations from automated data. The number of folds for combination one is three-folds, combination two is two-folds, and combination three is three-folds.

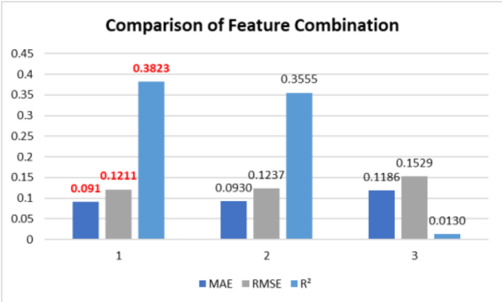


Figure 73. Comparison of TCT models from automated data using MLR

The result shows MLR obtains the lowest MAE value with 0.091, the lowest RMSE value with 0.1211, and the highest R^2 value with 0.3423 using combination one. It indicates that MLR develops a TCT model which obtains the lowest average error, close to most data points, and can capture most data points with combination two.

Support Vector Regression

Figure 74 shows the comparison between TCT predictive models that use the Support Vector Machine (SVR) method and different feature combinations from automated data. The number of folds for combination one is ten-folds, combination two is six-folds, and combination three is seven-folds.

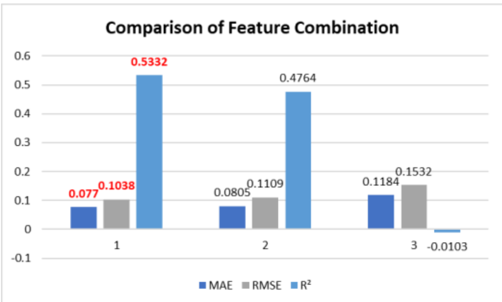


Figure 74. Comparison of TCT models from automated data using SVR

The result presents that SVR obtains the lowest MAE value with 0.077, RMSE value with 0.1038, and the highest R^2 value with 0.5332 using combination one. It indicates that SVR develops a TCT model which obtains the lowest average error, close to most data points, and can capture most data points with combination one.

Artificial Neural Network

Figure 75 shows the comparison between TCT predictive models that use Artificial Neural Network (ANN) method and different feature combinations from automated data. Combination one is developed by applying ten batch size, 500 epochs, 36 neurons in three hidden layer, and ten-folds. Combination two is developed by applying the ten batch size, 500 epochs, 36 neurons in three hidden layers, and four-folds. Finally, combination three is developed by applying the ten batch bize, 100 epochs, nine neurons in one hidden layer, and nine folds.

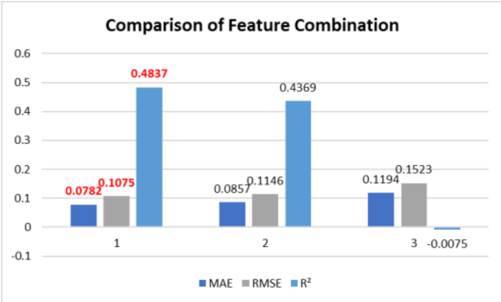


Figure 75. Comparison of TCT models from automated data using ANN

The result presents that ANN obtains the lowest MAE value with 0.0782, RMSE value of 0.1075, and the highest R^2 value with 0.4837 using combination one. It indicates that ANN develops a TCT model which obtains the lowest average error, close to most data points, and can capture most data points with combination one.

Overview

Figure 68 shows the comparison of TCT models from different methods in each combination with automated data. Combining one obtains the best predictive model with SVR, with the lowest MAE, RMSE, and the highest R^2 value. It indicates that combination one is suitable to develop a regression model using the SVR method. It is not far from the data points, has the lowest average error, and can capture most data points. With combination two, SVR has the lowest MAE, RMSE value, and the highest R^2 value. It indicates that combination two is suitable to develop a regression model using SVR. It can capture most data points, has the lowest average error. With combination three, ANN has the lowest RMSE value, SVR has the lowest MAE value, and MLR has the highest R^2 value. It indicates that combination three has a trade-off using different methods because each model has its strength.

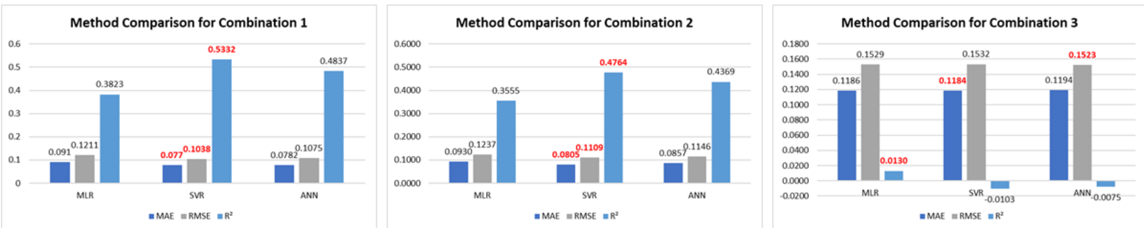


Figure 76. Comparison of TCT models from automated data

Overall, the best TCT model is developed by SVR with feature combination one from automated data. It has a similar performance with the predictive model from ANN with combination one.

4.6.3. Overview of Automated data

The predictive model using automated data has a better result than the predictive model using manual data. Table 8 presented the two best predictive model for each target using automated data. Most models are developed by ANN or SVR, except the TTT model that MLR develops. It also presents that most methods are developed using feature combination one or 2. The result indicates that feature combination three cannot develop a good predictive model by the selected methods.

Table 8. Overview result of automated data

Target	Method	Combination	MAE	RMSE	R^2
LT	SVR	1	0.068	0.096	0.362

	ANN	1	0.063	0.0981	0.3325
HT	SVR	2	0.068	0.089	0.464
	ANN	1	0.0670	0.0904	0.4518
UT	ANN	2	0.0291	0.0589	0.1340
	ANN	1	0.0303	0.0589	0.1118
RT	ANN	2	0.0429	0.0623	0.7203
	SVR	2	0.046	0.064	0.707
TTT	ANN	1	0.0511	0.0689	0.7485
	MLR	1	0.052	0.069	0.747
TCT	SVR	1	0.077	0.1038	0.5332
	ANN	1	0.0782	0.1075	0.4837

It presents that each model has a different average error, distance to data points, and ability to capture data points. It also shows that the UT model has the lowest average error and the closest model to the data points. However, it cannot accurately capture most data points, approximately 13% accurate. On the other hand, the TTT model is the best model to capture most data points which are approximately 74% accurate. However, high accuracy might cause overfitting, which is unable to be detected using k-fold cross-validation. Hence, the following section will explain the evaluation of each model and scenario using the test dataset.

5. Result

In the previous chapter, manual data and automated data are used for developing predictive models. The models have been developed using different combination features as the input and process with MLR, SVR, and ANN. The models are developed by concerning the error value, robustness, and computational time. This chapter aims to evaluate the performance of the selected predictive model from automated data. Therefore, each model will be tested using a test dataset that is different from the training dataset. The result will be denormalized to the original scale for evaluating the model in the original unit. Feature ablation will be used to find the contribution of each feature in each predictive model.

5.1. Denormalization

In the previous section, predictive models are developed with the normalized value. The result is difficult to be understood because of the range between zero and one. Hence, the result needs to be denormalized to the original range. Denormalization is the inversion from normalization value that aims to understand and evaluate the prediction result in the original range. Equation 19 is the denormalization equation which is indicated by x_i. The normalized value is indicated by z_i. The maximum value and the minimum value are indicated by $\max(x)$ and $\min(x)$, respectively.

Equation 19. Denormalization

$$x_i = z_i * \left(\max(x) - \min(x)\right) + \min(x)$$

Table 9 is the sample of denormalization result from the LT model. Test and Test Denorm refer to the ground truth in normalized and denormalized value, respectively. Predict and Predict Denorm refer to the prediction result in normalized and denormalized value, respectively. Deviation value refers to the different value between prediction and test in the original range and unit in seconds.

Table 9. Sample of Denormalization Result

Test	Predict	Test Denorm (Sec)	Predict Denorm (Sec)	Deviation Value (Sec)
0.1490	0.2108	385	435	50.2
0.1539	0.3649	389	560	171.3
0.1589	0.3898	393	581	187.5

5.2. Feature Ablation

The contribution of each feature to the predictive model will be evaluated by removing one feature and keep the rest features [29]. The objective is to identify that the performance of a predictive model is affected unequally by a particular feature. The predictive model which is affected by a certain feature will be analyzed further. Feature ablation will be conducted to the predictive model that has a good performance in the test dataset.

5.3. Model Evaluation

Each target has two predictive models, which shows a good result with the training dataset. The performance of a predictive model is tested using a test dataset that consists of 118 data points. The main part of the model evaluation is the robustness, prediction tendency, and feature contribution for the model. The robustness of the model will be evaluated based on the test result represented by the MAE value, RMSE value and R^2 value. The prediction tendency is assessed by analyzing the deviation value in the form of a box plot. The weakness of the model also is identified by analyzing the data points of outliers. It will help to understand which input is difficult to be estimated and the result of estimation. Table 10 shows the statistic data of each variable. This value will be used for analyzing the poor prediction results, especially the 25% and 75%, which refers to the interquartile range in a box plot. The feature contribution and importance will be assessed using feature ablation.

Table 10. Mean, Max, and Max value of each variable

	Distance (km)	Temperature (°C)	Relative Humidity (%)	Start Time Hour	Model	Volume (m^3)
Mean	1.66	15.9	73.27	12.35	0.2	17.15
Min	0.6	8.9	49.3	7	0	1.14
Max	3.9	18.2	95.4	17	1	27.7
25%	1.5	15.4	56.48	10	0	14.87
50%	1.66	16.1	79.43	12	0	17.5
75%	1.8	16.9	86.47	15	1	19.34

5.3.1. Load Time (LT)

Table 11 presents the test result of two LT models from SVR with combination one and ANN with combination one. The models obtain good performance indicated by a lower MAE and RMSE value and a higher R^2 value of the test dataset than the training dataset. The second model has better performance than the first model in predicting the test dataset.

Table 11. Evaluation of LT models

No	Method	Combination	Data	MAE	RMSE	R^2
1	SVR	1	Train	0.068	0.096	0.362
			Test	0.0663	0.0844	0.3910
2	ANN	1	Train	0.063	0.0981	0.3325
			Test	0.0569	0.0823	0.4219

Figure 77 shows box plots of deviation value after the result of LT models is denormalized. The red box plot represents the LT model with SVR and combination two, and the blue box plot represents the LT model with ANN and combination one. The red box plot has a wider interquartile range and more outliers below -100 seconds than the blue box plot. However, the blue box plot has more outliers above 100 seconds than the red box plot. It indicates that the first predictive model tends to overestimate the load time while the second predictive tends to underestimate the load time.

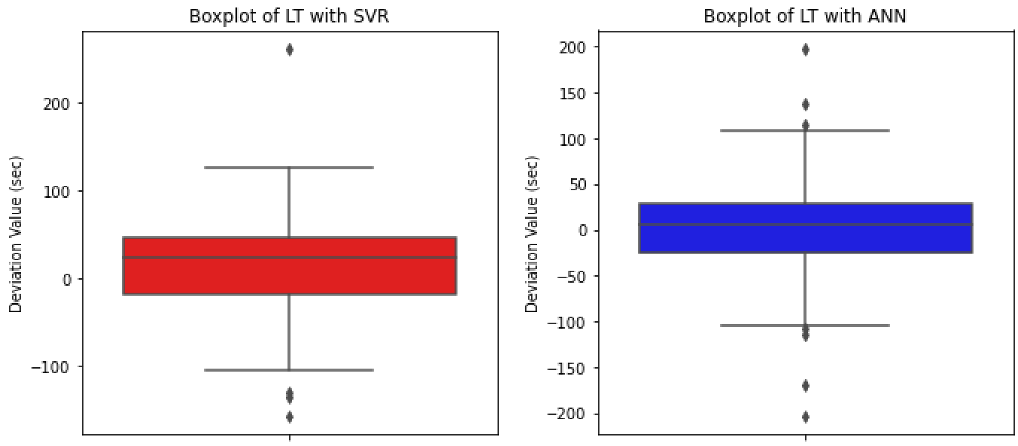


Figure 77. Box plot of LT models

The second model will be analyzed further, particularly its outliers. Table 12 shows the outliers data points, the prediction result, and the deviation values from the second LT model. The predictive model cannot accurately predict data points with the volume input is 14.7 m³ and 1.1 m³, and model input is one or Volvo A45G. Also, temperature and distance input values far from their median values in automated data are difficult to predict.

Table 12. LT model with ANN and combination one

Model	Start Time Hour	Volume (m³)	Relative Humidity (%)	Distance (Km)	Temperature (°C)	Test Denorm (sec)	Predict Denorm (sec)	Deviation Value (sec)
1.0	9.0	14.692982	80.96	1.947	12.7	344.0	140.403961	-203.596039
1.0	9.0	14.692982	82.24	0.957	12.4	383.0	213.683243	-169.316757
1.0	16.0	14.692982	49.64	0.778	17.7	326.0	210.810318	-115.189682
1.0	15.0	1.140351	50.71	1.715	17.9	107.0	243.624298	136.624298
1.0	16.0	14.692982	50.44	0.710	17.7	1.0	198.558807	197.558807

Figure 78 shows the result of the feature ablation in the LT Model with ANN and combination one. Distance, temperature, and model are identified to have a contribution to the predictive model. The predictive model obtains worse performance by eliminating those features. However, the result achieves better performance when volume and relative humidity are eliminated from the feature input. Thus, it indicates the model might have a better performance without having one of them.

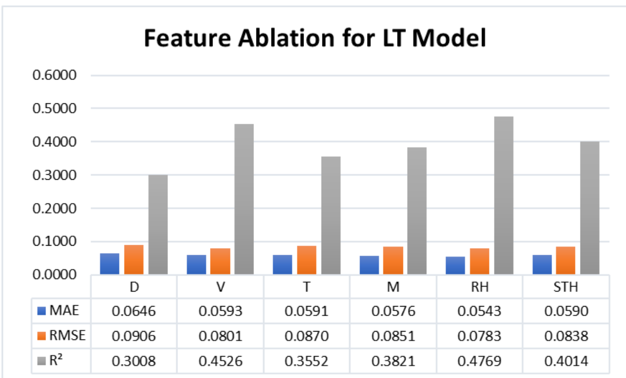


Figure 78. Feature ablation for LT model

The result indicates that the LT model with ANN and combination one has a robust performance in the test dataset. The most prediction has a deviation value between -50 to 50 seconds. In addition, the accuracy from the LT model is not high enough because it only achieves 42% accuracy. LT model is also required to keep distance, temperature, model, and start time hour as the input to ensure its performance. Volume and relative humidity might be considered to be removed from the input for better performance.

5.3.2. Haul Time (HT)

Table 13 shows the test result of two HT models from SVR with combination two and ANN with combination one. The models obtain poor performance indicated by a higher MAE and RMSE value and a lower R^2 value of the test dataset than the training dataset. The second model performs better than the first model in the test dataset, although it gave an opposite result with the training dataset. It shows that the second predictive model has better fitting and more robust than the first predictive model.

Table 13. Evaluation of HT models

No	Method	Combination	Data	MAE	RMSE	R^2
1	SVR	2	Train	0.068	0.089	0.464
			Test	0.0737	0.1121	0.3642
2	ANN	1	Train	0.0670	0.0904	0.4518
			Test	0.0700	0.1116	0.3701

Figure 79 shows box plots of deviation value after the result of HT models is denormalized. The red box plot represents the HT model with SVR and combination two, and the blue box plot represents the HT model with ANN and combination one. Both box plots have a similar interquartile range between -25 and 25 seconds and outliers below -50 seconds. The red box plot has more outliers above 50 seconds. The blue box plot shows better performance because it can predict more accurately and less error than the red box plot.

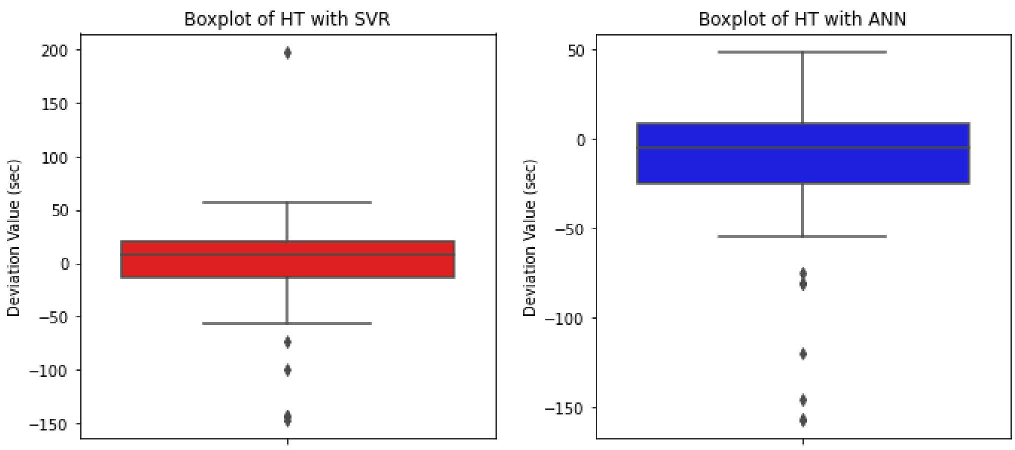


Figure 79. Box plot of HT models

However, the blue box plot indicates that the HT model overestimates some predictions. Table 14 shows the outliers data points, the prediction result, and the deviation values from the predictive model using ANN and combination one. The two highest deviation values have similar distance input, which is 1.62 km. The model might not predict well with the distance input is 1.62 km because it is the median value of the distance variable. It indicates many value input for that value but might have a different result because of other variables not counted in this research.

Table 14. HT model with ANN method and combination one

Model	Start Time Hour	Volume (m³)	Relative Humidity (%)	Distance (Km)	Temperature (°C)	Test Denorm (sec)	Predict Denorm (sec)	Deviation Value (sec)
1.0	10.0	20.745614	85.43	1.622	15.6	312.0	154.747711	-157.252289
0.0	8.0	10.964912	94.21	1.200	11.6	289.0	132.838135	-156.161865
1.0	16.0	18.245614	92.97	1.629	16.5	311.0	164.977982	-146.022018
0.0	15.0	19.868421	91.83	1.9	16.5	155.0	201.388382	46.388382
0.0	10.0	16.403509	67.77	1.9	14.4	165.0	213.616257	48.616257

Figure 80 shows the result of the feature ablation in the HT model with ANN and combination one. It shows that all features contribute to the model because the result has more error and less accurate when eliminating variables. The most contributing variable is the variable distance, where the HT model has a higher error because the variable distance is eliminated.

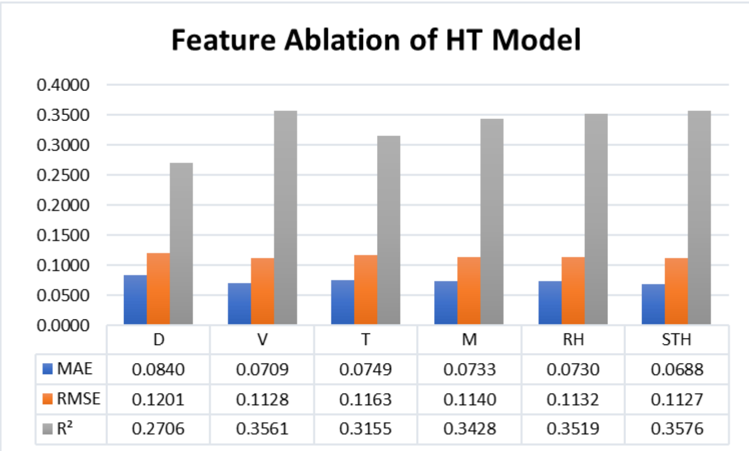


Figure 80. Feature ablation for HT model

The result indicates that the HT model with ANN and combination one performs better than the model with SVR and combination two. Most predictions have a deviation value between -25 to 25 seconds and tend to overestimate the hauling time. In addition, the accuracy from the HT model is not high enough because it only achieves 45% accuracy. The model is required to keep all the variables, particularly distance, to ensure the model performance.

5.3.3. Unload Time (UT)

Table 15 shows the test result of two UT models from ANN with combination two and ANN with combination one. The models have an overfitted pattern and obtain poor performance indicated by a higher MAE and RMSE value and a lower R^2 value of the test dataset than the training dataset. The first model performs better than the second model in the test dataset because the model has less error and more robust.

Table 15. Evaluation of UT models

No	Method	Combination	Data	MAE	RMSE	R^2
1	ANN	2	Train	0.0291	0.0589	0.1340
			Test	0.0349	0.0664	0.0574
2	ANN	1	Train	0.0303	0.0589	0.1118
			Test	0.0337	0.0669	0.0445

Figure 81 shows box plots of deviation value after the result of UT models is denormalized. The red box plot represents the UT model with ANN and combination two, and the blue box plot represents the UT model with ANN and combination one. Both box plots have a similar interquartile range around zero and outliers. The outliers led the models to have low accuracy. The blue box plot has more outliers than the red box plot. The red box plot shows better performance because it can predict more accurately and less error than the blue box plot.

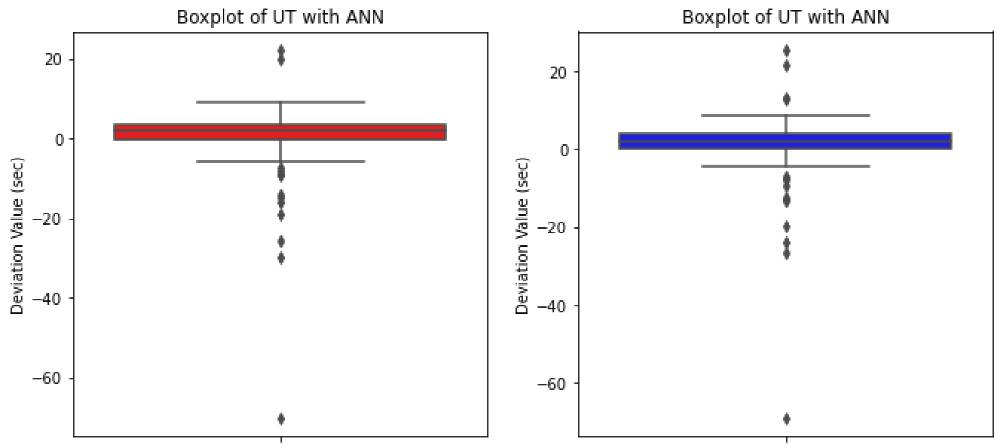


Figure 81. Box plot of UT model

The red box plot indicates that the UT model tends to overestimate some predictions. Table 16 shows the outliers data points, the prediction result, and the deviation values from the predictive model using ANN and combination two. The model might not predict well with the volume input is 14.7 m³, and model input is 0.7 km. It indicates that other variables that are not counted in this research affect the result.

Table 16. UT model with ANN and combination two

Model	Volume (m³)	Distance (Km)	Test Denorm (sec)	Predict Denorm (sec)	Deviation Value (sec)
0.0	18.026316	1.700	101.0	30.837704	-70.162296
1.0	14.692982	0.778	56.0	26.133913	-29.866087
1.0	14.692982	0.937	52.0	26.133913	-25.866087
0.0	10.394737	1.700	11.0	30.832125	19.832125
1.0	14.692982	0.710	4.0	26.133913	22.133913

Figure 82 shows the result of the feature ablation in the UT model with ANN and combination two. It shows that all features contribute to the model because the result is less when eliminating any variables. The variables significantly contribute to the predictive model because the error is higher and less accurate by eliminating distance.

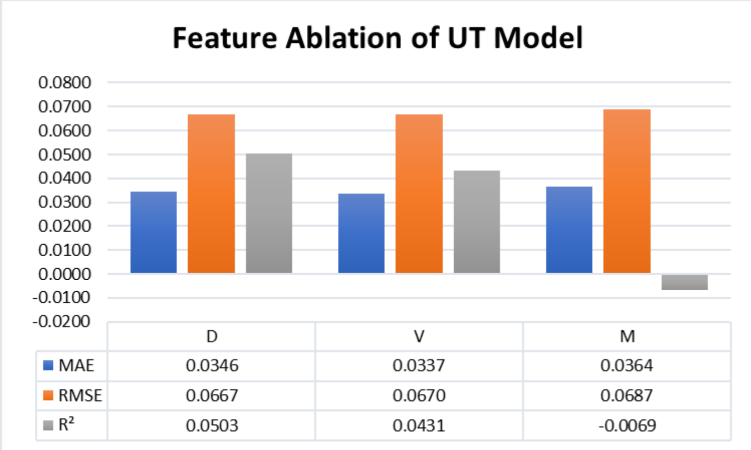


Figure 82. Feature ablation for LT model

The result indicates that the UT model with ANN and combination two performs better than ANN and combination one. Most predictions have a deviation value around zero seconds, but it tends to overestimate the unloading time. In addition, the accuracy from the UT model is not high enough because it only achieves 6% accuracy. The model has to keep all the input variables, particularly the variable model, to ensure the consistency of model performance.

5.3.4. Return Time (RT)

Table 17 shows the test result of two RT models from ANN with combination two and SVR with combination two. The models are robust, indicated by the higher R^2 value. However, it has more error indicated by a slightly increase MAE and RMSE value from the test dataset to the training dataset. It also presents that the first model performs better than the second model in the test dataset.

Table 17. Evaluation of RT models

No	Method	Combination	Data	MAE	RMSE	R^2
1	ANN	2	Train	0.0429	0.0623	0.7203
			Test	0.0498	0.0668	0.76
2	SVR	2	Train	0.046	0.064	0.707
			Test	0.0513	0.0672	0.7572

Figure 83 shows box plots of deviation value after the result of RT models is denormalized. The red box plot represents the RT model with ANN and combination two, and the blue box plot represents the RT model with SVR and combination two. The box plots have a different interquartile range where the red box plot range is between -25 and 25 seconds and the blue box plot range is between zero and 50. The red box plot shows better results than the blue box plot because the median value is close to zero, indicating that most of the prediction value is close to the actual value. The red box plot has more outliers above 50 seconds but fewer outliers below -70 than the blue box plot.

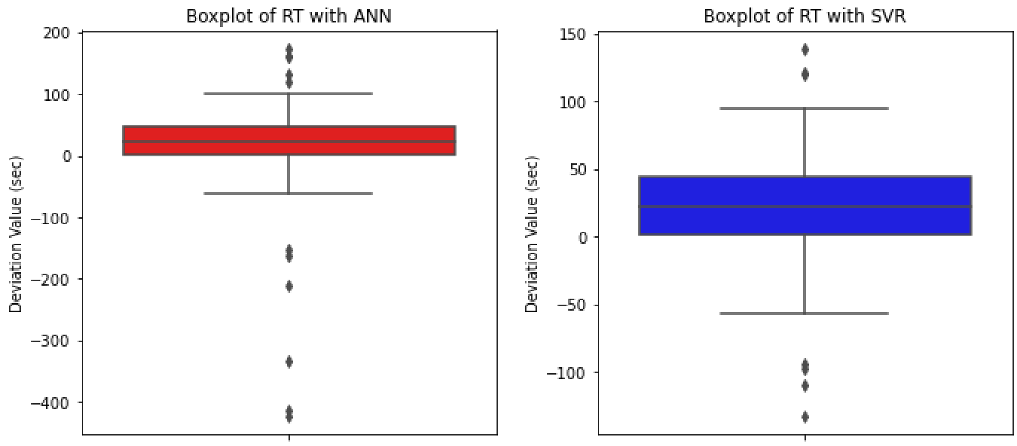


Figure 83. Box plot of RT models

However, the red box plot indicates that the RT model has tendencies to underestimate and overestimate some predictions. Table 18 shows the outliers data points, the prediction result, and the deviation values from the predictive model using ANN and combination two. The model might overestimate the return time if the value input is outside the interquartile range of each variable. And it might underestimate the return time if the distance input is around 1.6 km and the model input is 1. It indicates that other variables that are not counted in this research affect the result.

Table 18. RT model with ANN and combination two

Model	Volume (m^3)	Distance (Km)	Test Denorm (sec)	Predict Denorm (sec)	Deviation Value (sec)
0.0	21.403509	3.5	691.0	267.019012	-423.980988
0.0	12.938596	3.2	657.0	241.487503	-415.512497
0.0	17.587719	3.2	587.0	251.702713	-335.297287
1.0	18.245614	1.629	52.0	213.862961	161.862961
1.0	20.745614	1.622	46.0	219.194290	173.194290

Figure 84 shows the result of the feature ablation in the RT model with ANN and combination two. It shows that the variable distance significantly contributes to the predictive model because the error value is higher and less accurate by eliminating it. However, the model has a better performance by eliminating variable volume or model from the feature.

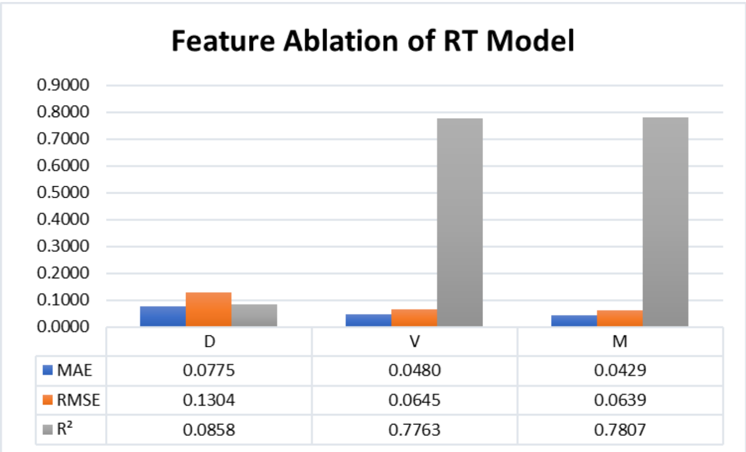


Figure 84. Feature ablation of RT model

The result indicates that the RT model with ANN and combination two performs better than the model with SVR and combination two. Most predictions have a deviation value between -25 to 25 seconds and tend to overestimate and underestimate the returning time. RT model has good performance because it only achieves 76% accuracy. The model is required to keep the variable distance to obtain a good prediction.

5.3.5. Truck Travel Time (TTT)

Table 19 shows the test result of two TTT models from ANN with combination one and MLR with combination one. The models obtain good performance indicated by a lower MAE and RMSE value and a higher R^2 value of the test dataset than the training dataset. The second model has better performance and shorter computational time than the first model in predicting the test dataset.

Table 19. Evaluation of TTT models

No	Method	Combination	Data	MAE	RMSE	R^2
1	ANN	1	Train	0.0511	0.0689	0.7485
			Test	0.0526	0.0673	0.7879
2	MLR	1	Train	0.052	0.069	0.747
			Test	0.0524	0.0666	0.7923

Figure 85 shows box plots of deviation value after the result of TTT models is denormalized. The red box plot represents the TTT model with ANN and combination one, and the blue box plot represents the TTT model with MLR and combination one. Both box plots have some outliers below -100 seconds. The red box plot has a narrower interquartile range and has more outliers above 100 seconds than the blue box plot. The blue box plot has better performance than the red box plot.

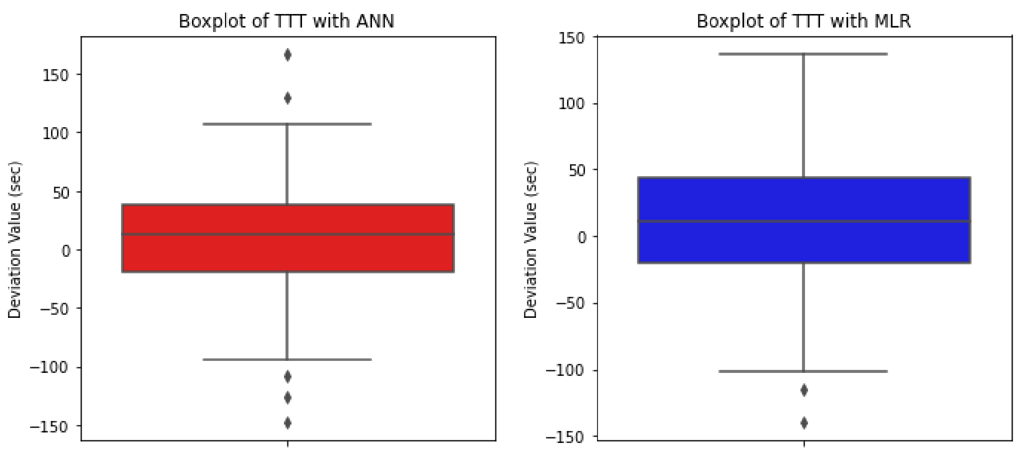


Figure 85. Box plot of TTT Models

However, the red box plot tends to overestimate the truck travel time for some predictions. Table 20 shows the outliers data points, the prediction result, and the deviation values from the predictive model using MLR and combination one. The model might overestimate the truck travel time if the distance input is outside the interquartile range of its variable and model input is model 0. It indicates that the model requires more historical distance data to develop a more robust model.

Table 20. TTT model with MLR and combination one

Start Time Hour	Model	Volume (m³)	Relative Humidity (%)	Distance (Km)	Temperature (°C)	Test Denorm (sec)	Predict Denorm (sec)	Deviation Value (sec)
11.0	0.0	17.500000	60.72	2.0	15.8	624.0	484.382044	-139.617956
12.0	0.0	16.622807	56.48	1.9	17.2	579.0	464.127118	-114.872882
16.0	0.0	21.359649	92.44	2.0	16.4	535.0	433.198074	-101.801926
9.0	0.0	15.526316	75.91	1.600	13.7	292.0	403.829528	111.829528
9.0	1.0	14.692982	80.96	1.947	12.7	309.0	445.615671	136.615671

Figure 86 shows the result of the feature ablation in the TTT model with MLR and combination one. It shows that the variable distance significantly contributes to the predictive model because the error value is higher and less accurate by eliminating it. However, the model has an equal and better performance by eliminating variable temperature, relative humidity, and start time hour from the input.

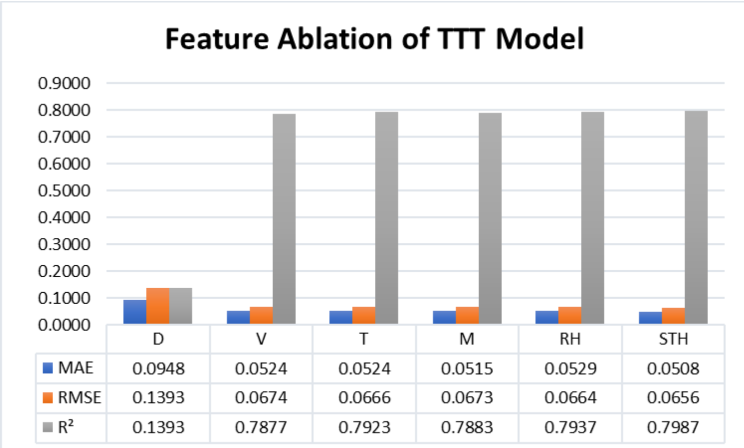


Figure 86.Feature ablation for TTT model

The result indicates that the TTT model with MLR and combination one performs better than the model with ANN and combination one. Most predictions have a deviation value between -25 to 50 seconds and tend to overestimate and underestimate the truck travel time. TTT model has good performance because it achieves 79% accuracy. The model is required to keep the variable distance for maintaining the performance.

5.3.6. Truck Cycle Time (TCT)

Table 21 shows the test result of two TCT models from SVR with combination one and ANN with combination one. The models obtain good performance indicated by a lower MAE and RMSE value and a higher R^2 value of the test dataset than the training dataset. In addition, the second model has better performance than the first model in predicting the test dataset.

Table 21. Evaluation of TCT models

No	Method	Combination	Data	MAE	RMSE	R^2
1	SVR	1	Train	0.077	0.1038	0.5332
			Test	0.0776	0.0987	0.5453
2	ANN	1	Train	0.0782	0.1075	0.4837
			Test	0.0692	0.0955	0.5740

Figure 87 presents box plots of deviation value after the result from TCT models is denormalized. The red box plot represents the TCT model with SVR and combination one, and the blue box plot represents the TCT model with ANN and combination one. The box plots have a similar interquartile range which is between -20 and 80 seconds. However, the blue box plot shows better results than the red box plot because the median value is close to zero, indicating that most of the prediction value is close to the actual value. In addition, the blue box plot has more outliers above 150 seconds and fewer outliers below -150 than the red box plot.

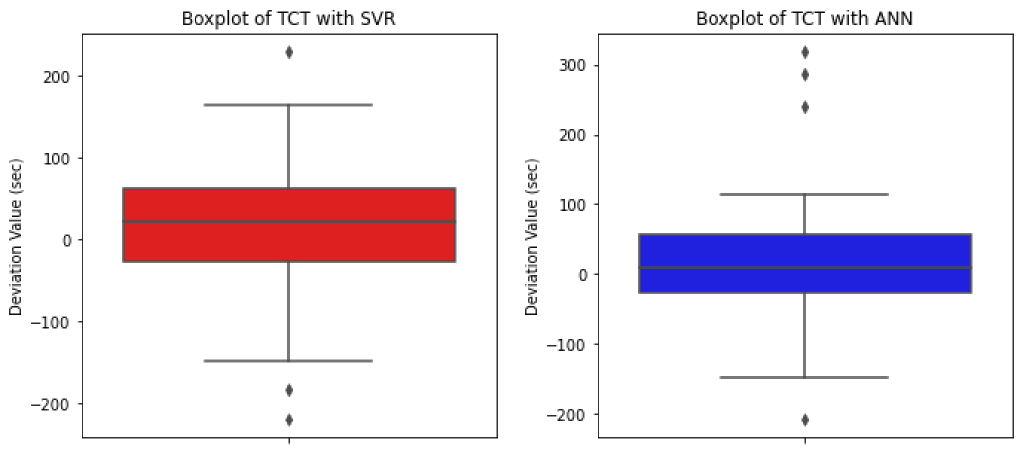


Figure 87. Box plot of TCT models

However, the blue box plot shows that the model overestimates and underestimates the truck cycle time for some predictions. Table 22 shows the outliers data points, the prediction result, and the deviation values from the predictive model using ANN and combination one. The model might not predict the truck cycle time if the input is outside the interquartile range of each variable. It indicates that the model requires more historical distance data to develop a more robust model.

Table 22. TCT model with ANN and combination one

Model	Start Time Hour	Volume (m³)	Relative Humidity (%)	Distance (Km)	Temperature (°C)	Test Denorm (sec)	Predict Denorm (sec)	Deviation Value (sec)
0.0	11.0	12.938596	63.63	3.200	16.1	1073.0	864.516296	-208.483704
1.0	9.0	14.692982	82.24	0.957	12.4	651.0	501.140198	-149.859802
0.0	17.0	10.394737	52.29	1.700	17.1	393.0	633.343140	240.343140
1.0	16.0	14.692982	50.44	0.710	17.7	264.0	551.515747	287.515747
1.0	15.0	1.140351	50.71	1.715	17.9	468.0	786.846191	318.846191

Figure 88 shows the result of the feature ablation in the TCT model with ANN and combination one. It shows that the variable distance significantly contributes to the predictive model because the error is higher and less accurate by eliminating it. However, the model has a better performance by eliminating variable relative humidity from the input.

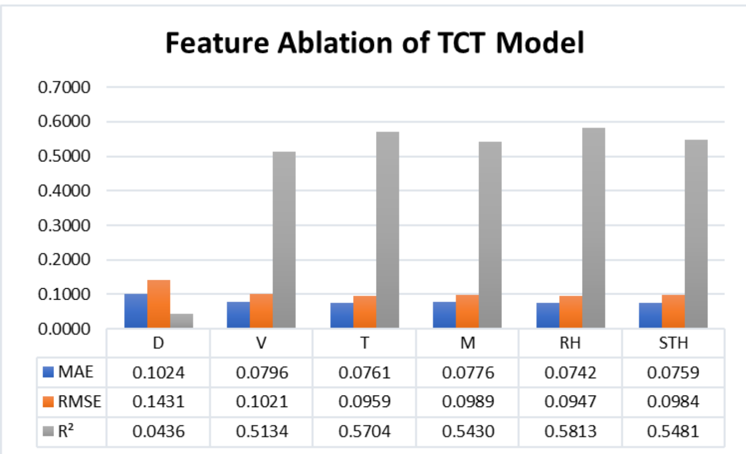


Figure 88. Feature ablation for TCT model

The result indicates that the TCT model with ANN and combination one performs better than the model with SVR and combination one. Most predictions have a deviation value between -25 to 50 seconds and tend to overestimate and underestimate the truck cycle time. TCT model has good performance because it achieves 57% accuracy. The model is required to keep the variable distance for maintaining the performance.

5.3.7. Overview of Model Evaluation

Table 23 shows the overall performance evaluation for the two best models for each target from automated data. The first row is the best model, and the second row is the second-best model of each target. It also shows the overview of MAE, RMSE, and R^2 value with train and test dataset. It shows that LT, RT, TTT, and TCT models have better performance in the test dataset than the training dataset, but the other models have worse performance in the test dataset than the training dataset. They have different performance result where the best model is the RT model with 79% accuracy, and the worst model is UT with 4.5 % accuracy.

Table 23. Predictive models evaluation

Target	Regression Technique	Combination	MAE		RMSE		R^2	
			Train	Test	Train	Test	Train	Test
LT	ANN	1	0.063	0.0569	0.0981	0.0823	0.3325	0.4219
	SVR	1	0.068	0.0663	0.096	0.0844	0.362	0.391
HT	ANN	1	0.067	0.07	0.0904	0.1116	0.4518	0.3701
	SVR	2	0.068	0.0737	0.089	0.1121	0.464	0.3642
UT	ANN	2	0.0291	0.0349	0.0589	0.0664	0.1340	0.0574
	ANN	1	0.0303	0.0337	0.0589	0.0669	0.1118	0.0445
RT	ANN	2	0.0429	0.0498	0.0623	0.0668	0.7203	0.76
	SVR	2	0.046	0.0513	0.064	0.0672	0.707	0.7572
TTT	MLR	1	0.052	0.0524	0.069	0.0666	0.747	0.7923
	ANN	1	0.0511	0.0526	0.0689	0.0673	0.7485	0.7879
TCT	ANN	1	0.0782	0.0692	0.1075	0.0955	0.4837	0.574
	SVR	1	0.077	0.0776	0.1038	0.0987	0.5332	0.5453

Most targets have ANN as the best and SVR as their second-best of regression techniques. Feature combination one, consisting of variable distance, start time hour, volume, relative humidity, temperature, and model and feature combination two, consists of distance, volume, and model input. It shows that ANN is more robust in predicting new input than the other regression models because the test dataset results are higher than other models. In addition, the result between models of each target has a similar value where the difference is not more than five per cent. The previous analysis of deviation value comparison with box plots shows the interquartile range difference is approximately 20 seconds. Hence, it can be an opportunity to use different models according to the user intention.

5.4. Evaluation of Regression Techniques

The evaluation of regression techniques aims to know the strength and weakness of the regression techniques in developing or using models. It is important to indicate that the evaluation result might differ if the regression techniques solve different problems. The evaluation of predictive models from different regression techniques can be concluded by considering the prediction accuracy, ease of development and use, and transparency. Table 24 shows the evaluation comparison result between MLR models, SVR models, and ANN models. The evaluation is given and comparatively determined with a high, medium, or low performance toward most models. For instance, most ANN models have a higher prediction accuracy compared with MLR and SVR models.

Table 24. Evaluation of regression techniques

Evaluation Aspect	Model		
	MLR	SVR	ANN
Prediction accuracy	Low	Medium	High
Ease of development and use	High	High	Low
Transparency	High	Medium	Low

Prediction Accuracy

The prediction accuracy evaluation is derived based on the performance metrics of the models. The evaluation models with test dataset show that ANN models have a higher accuracy compare to other models. And the SVR models have higher accuracy than the MLR models. The regression technique ability of the non-linear modelling relationship influences the model's accuracy because the relation between some variables is non-linear. MLR can only develop a linear relationship, while SVR and ANN can develop a non-linear relationship. However, the best regression techniques for predicting TTT is MLR which has the highest accuracy compared to other models. It indicates that the model's accuracy also influenced by the provided data, not solely based on the ability of the regression technique.

Ease of Development

Besides predictive models accuracy, the ease of model development and use must be evaluated as the user consideration in using the model. The ease of development is derived based on the hyperparameter tuning and testing process of the models. MLR and SVR are easier to develop and used because both techniques are less complex than ANN. The number of parameters in ANN is more than others, so it requires a high iterative process to develop the models. The ease of use is derived from the computational time to process the input and give the output. ANN has the longest time processing the algorithm than MLR and SVR because the algorithm's complexity causes the computational time to be longer. The data size also affects the ability of regression techniques to process it. MLR and SVR are faster in processing small data but slower in processing big data compare to ANN. Therefore, ANN might not use a good regression technique if the user requires a short time to predict TCT.

Transparency

Transparency of the model outcome and the development process is evaluated because it relates to the trust toward the engineers who will develop, maintaining, and updating the model. The transparency of the model is derived based on the accessibility to know the final mathematical model. MLR is easier to interpret than SVR and ANN because the coefficient of each variable and the model's intercept can be obtained (see Appendix 3). SVR can be visualized to understand how the hyperplane works in developing a regression model. However, SVR is difficult to be visualized if the input has many variables. ANN is not a transparent regression technique because how the variables reach the final prediction is complex and difficult to understand [30]. Therefore, the transparency and interpretability of MLR can help the user to understand the previous method's weakness and improve it than other regression techniques.

Overview

The evaluation of regression techniques in developing predictive models indicates the trade-off in using a predictive model. MLR is easy to use and develop a predictive model and transparent in understanding the mathematical model. However, it is only good for predicting the linear relationship, not the non-linear relationship. Hence, it is suitable for a user who aims to understand mathematical modelling and predict a linear model, such as the TTT model.

SVR is easy to use and developed and can develop non-linear relationships, but the mathematical model's transparency is difficult to obtain. Hence, it is suitable for a user who aims to obtain a fast result from a non-linear model with small data size. ANN can develop a non-linear relationship with high accuracy and robustness, lower range of deviation value. It can handle big data, but it has a high complexity in developing and understanding the mathematical model. ANN might be an overkill regression technique for processing a small data size. Hence, it is suitable for user to aims to obtain an accurate prediction.

5.5. Scenario Evaluation

Figure 89 shows box plots of deviation value from different scenario to predict TCT. It shows that the first scenario, the sum of the individual model, has the widest interquartile range, the most outliers, especially above 200 seconds. The second scenario is the sum of the LT, UT, and TTT model. It has a smaller interquartile range than the first scenario, which is approximately 6 seconds or 10 per cent decreased. It also has the least outliers among other scenarios. Lastly, the third scenario, the TCT model, has the smallest interquartile range but the most outliers below -150 seconds.

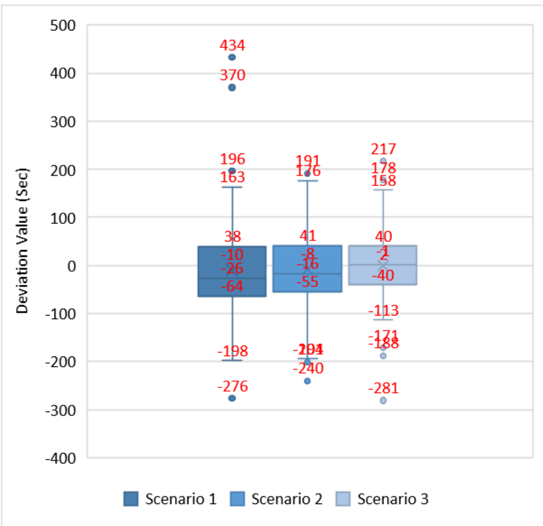


Figure 89. Box plot of TCT scenario

The result indicates that each scenario to obtain TCT has their strength and weakness. The first scenario is good for finding the detailed activity time. Experts can analyse the result, set a buffer time in each activity, or optimize a certain activity time. However, the first scenario has the least accuracy amongst all scenarios, and it requires more time to get the result of TCT. Moreover, it requires each model to give a result and accumulate the result. The second scenario is good for finding the detailed activity time of LT and UT related to different machinery types and site conditions. It also has the least outliers, which helps experts know the close value from the actual value. However, it requires to obtain LT, UT, and TTT and accumulate the result. And the individual time of HT and RT is unknown using this scenario. The third scenario is good for predicting TCT that closes to the actual value quickly because it has the highest accuracy in predicting TCT and requires only to run the TCT model to obtain the prediction value of TCT. However, the third scenario cannot optimize individual activity duration and might have underestimated or overestimated value if the input is outside the interquartile range. Therefore, each scenario can be useful in different ways depending on the aim of the prediction besides obtaining accurate time prediction.

6. Practical Implications

The previous chapters focus on the result of predictive models from automated data. This chapter evaluates the predictive models from a practical perspective and the implication in monetary benefit. Therefore, this chapter will explain the stakeholder perspective based on the interview, calculate the monetary benefit, and suggest using the predictive model.

6.1. Stakeholder Interview

An accurate truck cycle time prediction is important for stakeholders in earthworks, especially contractor and construction machinery company. The interview was conducted with experts to understand the finding from the predictive model and gain the perspective from the practical point of view. The interview method used is a semi-structured interview because this method can develop a new discussion or idea as the interviewee's response during the interview. Different stakeholders might have different idea or perspective because they have different interest and goal.

The interview's main structure consists of two main parts: the current condition and the predictive model. The first part aims to know the current condition in obtaining an accurate truck cycle time. The second part aims to know their opinion about the predictive models and the finding found in the data. The answer from the contractor and construction machinery company will be explained in the following section.

6.1.1. Contractor

A general contractor or contractor is an individual or organization that a client hires to execute the project by building, oversight the construction process, managing the resources and trades. In general, the contractor aims to gain profit by delivering the project. Client trust is an important aspect for them to running their work or business. Contractors usually use sub-subcontractors for earthworks projects because it is a large project which requires many resources.

A sub-contractor is a smaller contractor with a specialist for particular construction works employed by a general contractor. Sub-contractors calculate the estimation of truck cycle time because they know the resources better than a general contractor, which focuses more on managing the sub-contractors.

The interview was conducted with a general contractor's project manager who has experience managing sub-contractors and resources for earthworks and communicates with clients. The shared perspective from the project manager also considers other roles, for instance, driver, for the practical implication of this research.

Current Practice

The predictive model might be a useful tool to replace the current practice using simple math. The tool to estimate is important for the contractor because it helps contractor truck cycle time in earthwork. The contractor needs to set up a good strategy in allocating resources because it affects contractor performance, anticipation plan, and client trust. The predictive model has advantages in knowing the accuracy level, considering many variables, fast calculation, and fewer human resources estimating the TCT. The expected accuracy of the predictive model is approximately more than 80%. However, the uniqueness of the project and the unexpected aspect is considered to be difficult to expect 100% accuracy. The most important part is to set up an effective anticipation plan in delivering the project.

The Predictive Model

All the predictive model has a positive contribution in earthworks, especially the load time and unload time. Load time is important to set up the loader excavator to ensure the trucks do not need to wait. Unload time is also important because unloading is sometimes limited, and trucks need to wait. The predictive model is also useful for moving overburdened material, which usually has a short distance between the loading and unloading location.

The variable that is counted also covers more variables than the current practice. The most important variables are start time hour and distance variables, among the other features in this research. Start time hour as an input will help to manage the human resources and the machinery. Distance has a contribution to calculating the truck cycle. Usually, the starting point, load area, unload area, and endpoint are known before the earthworks are started.

However, the robustness of the predictive model from this research is not enough to be implemented in real work. The project manager also pointed out the important part of counting the operation practice, such as driver behaviour and operation condition factors, such as road conditions and weather conditions.

6.1.2. Construction Machinery Company

A construction machinery company or supplier is a manufacturer of construction machinery and equipment, for instance, trucks and excavators. They have a pivotal role in supplying machinery to the client. In earthworks, the client is the contractor who is the buyer or dealer who provides rental service for the machinery. They put safety as their priority in delivering their product. It also an important aspect of gaining client trust.

The estimation of truck cycle time indirectly benefits the company because the market demand depends on the strategy used by the contractor. As an instance, the contractor will order more trucks from the supplier to obtain their project goal. Therefore, the time estimation also gives a direct benefit in giving the information of truck performance.

The interview was conducted with a logistic and operation analyst at one of the world's leading construction machinery company. The interviewee has many experiences in analysing the good quality machinery production and procurement process of various earthworks machinery.

Current practice

The supplier has a confidential strategy to calculate the TCT by utilizing the data from the user in a certain time frame. Then, the engineering team will analyze the calculation result for improving the performance of their company. Hence, the supplier improvement depends on the given data from the contractor. Bad data will lead to bad improvement of the machinery and client distrust.

Moreover, the estimation of TCT affects the demand that their company suddenly requests. The requested machinery has a large size where the company has a limitation of warehouse to keep them. It impacts how the company use their resources and fulfil the request. Therefore, the predictive model helps the supplier to have reliable data from the company that uses it.

The Predictive Model

A positive response is given to the development of predictive models, especially to the load time. Loading activity involves more machinery, for instance, excavator and trucks, than the other activities in earthworks. The calculation in loading material is more complicated than the other activity because the type of soil, the combination of machinery, the condition of the site impact the result.

Model is counted as the most important variable in predicting the truck cycle time from the supplier perspective. The right machinery needs to be chosen in earthworks to deliver the project. Despite the limitation of predictive models, they can be used in most earthworks.

However, the robustness of the predictive model from this research is not enough to be implemented in real work. The supplier also pointed out the important part of counting the operation practice, such as driver behaviour and experience. A driver should have enough skill and experience in operating the machinery. The supplier highlighted the difficulty to operate the excavator for loading the material into the truck bucket. Moreover, the site condition in earthworks can be dangerous for all activities.

6.2. Cost and Benefit Analysis

Based on the stakeholder perspective and discussion with experts in the construction industries, the accuracy improvement obtained from predictive models might benefit the stakeholders. It might replace the current method or traditional method in predicting TCT. However, replacing the current method with the predictive models requires cost, for instance, operational cost. Therefore, the following section will analyse the benefit and cost of predictive models.

6.2.1. Benefit

This research will analyze the tangible and intangible benefits of the predictive models that are used in scenarios. Tangible benefits are benefits that can be quantified, for instance, fuel consumption. They are analyzed by calculating the deviation between TCT prediction and actual value from predictive models and the traditional method. The deviation value indicates the inaccurate value of TCT prediction, which causes a queue time of a truck to load or unload material. A truck's queue time is considered ineffective because the resources are wasted in that activity, for instance, fuel and human resources.

The traditional method will be used as the benchmark for the benefit comparison. The traditional method will use Equation 2 for calculating haul time and Equation 3 for calculating return time. Because of data and information limitation, the load and unload times will use the average duration for those activities, 30 seconds [8]. It also uses 54.8 km/hour for the truck speed, which is the average speed for truck models.

Intangible benefits are benefits that are difficult to be quantified, for instance, client trust. They are analyzed based on the discussion result from experts.

6.2.1.1. Tangible Benefits

This research will analyze two tangible benefits are considered by contractors, such as the accuracy of truck productivity and the monetary benefit. The accuracy of truck productivity is important for contractors to manage their resources such as equipment and workers. More accurate the truck productivity prediction, the better the contractor to manage the resources. The monetary benefit will be calculated from the fuel consumption and drivers productivity because they are two of the main expenses in the earthworks.

The calculation will use the comparison result of deviation value in predicting the test dataset. The test dataset consists of 118 data points or truck cycles, representing two truck activity in two days. Therefore, it assumed that one truck has approximately 30 cycles to transport the materials.

Truck Productivity Accuracy

An accurate TCT prediction has a benefit in increasing the accuracy of truck productivity prediction. Estimating truck productivity can help a contractor set up a strategy so the target can be achieved within the given time by a client, such as increasing the number of trucks. Table 25 shows the comparison between productivity accuracy results obtained from scenarios of TCT and the traditional method. The productivity is calculated by comparing the productivity prediction with the actual productivity value from the test dataset. Productivity is the calculation of TCT (hour) is divided by volume (m^3).

The first row shows the average inaccurate value of productivity prediction for two trucks in two days. The inaccurate value of productivity prediction will be higher along with the increasing number of trucks and operational days. The second row shows the comparison accuracy with the traditional method. It shows that the productivity accuracy using prediction models are approximately 20% is more accurate than the traditional method.

Table 25. Productivity accuracy comparison

Item	Scenario 1	Scenario 2	Scenario 3	Traditional Method
Average inaccurate value of productivity prediction (Hour/m^3)	10.12	10.61	10.96	46.94
Comparison of Accuracy with Traditional Method (%)	121.6	122.6	123.4	100

Monetary Benefit

Table 26 presents the comparison of the monetary benefit between scenarios of TCT and the traditional method. The first row shows the deviation prediction value from the actual value from different prediction methods. It shows that scenarios of TCT are approximately five times more accurate than the traditional method. Scenario three has the most accurate prediction.

The second row shows the inefficient fuel consumption, which is the multiplication result of inefficient TCT and the fuel consumption. The amount of fuel consumption that is used in the calculation is 12.11 litres/hour. It shows that scenarios of TCT have less waste fuel which is approximately five times than the traditional method. It will also reduce the amount of pollution which comes from burning fuel.

The third row shows the inefficient cost for fuel which is the multiplication of the average fuel consumption and the fuel cost. The fuel diesel cost that is used in the calculation is 1.5 € per liter. The fourth row shows the inefficient cost for drivers, resulting from the multiplication of driver salary per hours and the inefficient truck cycle time. The driver salary that is used in his calculation is 20€ per hour. The fifth row shows the total inefficient cost is the sum of the inefficient cost for fuel and drivers. It shows that accurate prediction benefits the contractor by reducing the inefficient cost approximately five times than the traditional method.

Table 26. Monetary benefit comparison for TCT

	Scenario 1	Scenario 2	Scenario 3	Traditional Method
Inefficient Truck Cycle Time (Hours)	2.30	1.98	1.83	11.89
Inefficient fuel consumption (Litres)	27.81	23.96	22.17	144.06
Inefficient cost for fuel (€)	41.71	35.93	33.26	216.10
Inefficient cost for drivers (€)	45.91	39.55	36.61	237.86
Total inefficient cost (€)	87.62	75.49	69.87	453.96

In addition, the predictive models of TTT have more benefits than the traditional method. Table 27 compares the sum of the HT and RT model, the TTT model, and the traditional method. In this analysis, the traditional method only includes haul time and return time. The result shows that the predictive model is approximately five to seven times more beneficial in reducing the inefficient cost for drivers and fuel.

Table 27. Monetary benefit comparison for TTT

	HT + RT	TTT	Traditional Method
Deviation with the actual duration (Hours)	1.74	1.27	8.94
Inefficient fuel consumption (Liters)	21.13	15.44	108.31
Inefficient Cost (€)	31.69	23.16	162.46
Inefficient cost for drivers (€)	34.89	25.5	178.82
Total inefficient cost (€)	66.58	48.66	341.28

6.2.1.2. Intangible Benefits

The predictive models have multiple intangible benefits: safety, trust in the decision-making process, worker satisfaction, and client trust. Better accuracy of TCT prediction will help to know the downtime of the machinery, road condition or routes. The monitoring process can have better planned to ensure safety in the construction projects. It also helps the sub-contractor gain the general contractor's trust, and the general contractor can check the correct estimation of TCT. This situation will help the general contractor to decide which sub-contractor that can be trusted. The accuracy also helps to obtain client trust. The worker's satisfaction might also increase because the machinery and human resources management can be managed better.

6.2.1.3. Overview

The practical implementation of the predictive model benefits different stakeholders, such as general contractor, sub-contractor, client, and machinery supplier. A Sub-contractor can gain general contractor trust and reduce unnecessary expense in fuel consumption, drivers, and machinery. It also helps to have a robust maintenance schedule of road and machinery. It is

also beneficial for the general contractor to decide which sub-contractor to work with based on the performance to manage their resources. A general contractor can increase client trust and better management of resources in completing the project. A general contractor also can obtain rewards because the project can finish earlier if the contract applied that system or avoid a fine because of the delay. A supplier can reduce the expense of overtime worker and increase the well-being of the employee. A client who owns the project will trust the contractor for completing the project.
More benefit will be gained by including other benefits, for instance, the environmental impact and the reduction of working hours to obtain TCT prediction. It also increases along with the robustness of the predictive models, where the accuracy of the TCT model from scenario three is only 57%. Hence, the predictive models can be used for improving the estimation accuracy of TCT in earthworks.

6.2.2. Cost

The predictive models require operational cost for paying engineers who will operate and maintain the predictive models, which approximately costs 27€ per hour, and paying the open weather service data, which costs approximately 28.71€ per month. The engineers also can help the reduction of inefficient work on the bigger scale of the contractors. It indicates that more work can be done in a short time. Hence, the predictive models bring more monetary benefit than the traditional method for earthworks. Moreover, both calculation only counts for approximately two-day in earthworks for two trucks. The monetary benefit will be increased by the increased number of trucks and days.

6.3. Strategy to Implement the Predictive Model

This research would like to suggest practical suggestions to the contractor, especially the planner, using the predictive model. This strategy Figure 90 is the main steps of the strategy in using the TCT model, which aims to improve the estimation of the truck cycle time.

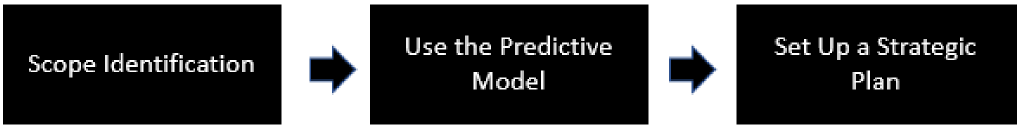


Figure 90. Strategy for using TCT model

1. Scope Identification

The similarity of the project needs to be should be analyzed before using the predictive model. The predictive model will not be reliable enough to predict TCT where the new input is different from automated data. The input limitation for each variable was explained in section 3.2.2, and the material type is limited to the overburden. TCT model from scenario three is recommended to be used for estimating TCT. TTT model is also recommended to be used for estimating TTT.

2. Use the predictive model

The predictive models are applied by inserting the input value for each variable. The variables input are distance, volume, relative humidity, temperature, start time hour, and model. The distance variable is the most required variable for the TCT model and TTT model. The models might obtain better performance by eliminating variable relative humidity from the input for the TCT model and eliminating variable temperature, relative humidity, and start time hour from the input for the TTT model.

3. Set Up a Strategic Plan.

An expert or engineer should analyse the prediction value because the predictive models do not have 100% accuracy. The estimation from TCT and TTT models should be analyzed if the input is outside the interquartile range of each variable *(see Table 10)*. The experts can set up the anticipation and strategic plans to manage the machinery and resources, such as setting up a buffer time or adding more trucks to complete the project.

7. Conclusion

This research has presented historical data for developing predictive models by utilizing various machine learning approaches to improve truck cycle time in earthworks. This research started with the identification of the problem gap by conducting a literature review. Then, truck cycle time and the affected factor of TCT in earthwork are also investigated. The affected factors are used as the starting point of which data need to be collected.

Data Exploration was conducted to understand the data quality and preparing the data to be an input for predictive models. The modelling process also has been explained where parameter tuning is conducted to find the best possible model. The predictive models also have been tested to know the robustness. Moreover, the practical implication of the predictive model for stakeholders has been explicated.

This chapter aims to conclude this research and give practical and scientific recommendations for improving the estimation of TCT in earthworks. This chapter will discuss the main finding and limitation of this research. The contribution and future recommendation for practical and scientific will be explained based on the research result.

7.1. Discussion

This section will present the answer to each sub-questions and the limitation of this research.

7.1.1. Research Questions Answers

This research started with a research question based on the problem gap and the opportunity to fill the gap. The following is the main research question of this research.
How can the historical data be utilized to improve the prediction accuracy of truck cycle time in earthworks?

The main research question is answered by addressed the sub-questions. The answer to each sub-question will be addressed as follows.

1. Which variables in the historical data should be included in the prediction model of truck cycle time in earthworks?

The first sub-question was initially answered through the literature review and data exploration. The literature review was intended to find the affected factor used as the starting point to collect historical data. This research uses manual data and automated data for earthmoving activity and weather data. Each data is explored for examining its quality and finding feature combination by conducting data exploration *(see chapter 3. Data Preparation)*. Manual data has four input variables and three feature combinations. Automated data has six input variables: distance, volume, relative humidity, temperature, start time hour, and model. The input variables are combined into three different feature combinations
However, the development and evaluation of predictive models found that input variables affect differently in each model. The variable contribution in each model is found through the feature ablation. In addition, the input value for each variable affects the prediction accuracy in each model.

2. How to develop an accurate predictive model of truck cycle time using the machine learning approach?

The second sub-question is answered by developing the predictive model from manual data and automated data. The manual data can develop the TCT model without RT. Automated data can develop the TCT model, TTT model, and individual activity model. The availability data of each activity duration opened the opportunity to develop different scenarios for predicting TCT (*see section 3.3. Scenario*). Different scenarios helped find the effective way to use the models to achieve the best estimation of TCT.

The regression method is chosen because it is a suitable method for continuous value. MLR, SVR, and ANN regression techniques are chosen based on the data preparation result. Hyperparameter tuning was conducted for obtaining the regression model. The most potential of a predictive model is listed in Table 8 based on the performance evaluation with the training dataset. The assessment of predictive models used error measurement using MAE and RMSE and the goodness of fit using R^2. The result shows that the historical automated data can develop the predictive model of TTT and RT with a good performance. The best predictive model for TTT uses MLR with feature combination one, and RT uses ANN with combination two.

The models were evaluated using the test dataset. Based on the evaluation outcome, the best model is chosen to predict TCT with different scenarios. The first scenario is the sum of individual prediction time from the LT, HT, UT, and RT model. The second scenario is the sum of truck travel time, load time, and unload time prediction from TTT, LT, and UT model. The third scenario is the TCT prediction from the TCT model. The result presents the best result obtained by scenario three or the TCT model with 57% accuracy.

3.What is the practical implication of using the predictive model of truck cycle time?

The answer to the third sub-question is explained in chapter 4. The contractor and machinery supplier is interested in having an accurate predictive model to improve their work. Predictive modelling has a direct benefit for the contractor and an indirect benefit for the machinery supplier. The predictive model might have a practical implication in setting up a plan before the project start. The plan consists of the management of people and machinery and the anticipation plan in delivering the project. The plan impacts the machinery supplier in managing their resources to fulfil the demand.

The predictive model might replace the traditional method in estimating the truck cycle time because it can consider many variables, which is difficult to be done in the manual calculation. It is also able to cover earthworks, particularly for transporting overburdened material. However, the predictive model is not reliable enough to be implemented because the accuracy is not good enough, and operation practice has not yet been included. It required experts to analyse the prediction from the predictive model.

The benefits of the predictive model are analyzed by comparing the scenarios and current practice or traditional method in predicting truck cycle time in the test dataset, which represented two trucks in two days. Based on the calculation, scenarios are approximately 20% more accurate in predicting truck productivity. Scenarios can also decrease inefficient truck cycle time approximately five to six times from the traditional method. The reduction of inefficient truck cycle time can reduce the cost for fuel and drivers, fuel emission. Based on the comparison result, stakeholders might gain tangible benefits, such as reduce inefficient expense. They might also gain intangible benefits, such as gaining partners trust, a better strategic plan to complete the project, and increasing the employee's well-being. Therefore, The prediction model has more benefits than the traditional method.

This research shows that historical data can be used for building a predictive model. Although not all predictive models have a high accuracy result, the predictive model has a better performance than the traditional method.

7.1.2. Limitation

This research has limitations in pursuing a better outcome, such as time constraint, the lack of project documentation, and limited resources. Time constraint in finishing this research affects the duration to obtain data. Hence, this research cannot produce high accuracy of predictive models because of the limited data. Time constraint also affects the model development. This research generated the weather data using the start time of a truck cycle. However, the weather might change in truck cycle time. For instance, the temperature might increase in hauling and decrease when unloading the material. This research did not conduct model exploration with different combination data between weather and operational time due to time constraint and the capacity of a computer to run the model. Hence, it also might impact the predictive model accuracy.

Lack of project documentation hindered the understanding of the data. The limited resources that refer to the limitation of people that can also be asked hindered data exploration. This limitation is handled using an assumption that should be considered in applying the predictive model.

7.2. Contribution

This research found some practical and scientific contributions in the process to answer the research questions. The following is the explanation of each contribution.

7.2.1. Practical

The practical contribution is found when the research has an assumption that the correlation between volume/weight and the target should be positive, not negative. However, stakeholders agreed with the negative correlation. The size of the machinery has different capacity and features, which makes the bigger excavator or bigger truck faster than the small machinery. However, a big excavator cost more money than a small excavator. It also depends on the site condition where the scattered material will take more time to be loaded. The contractor considers the trade-off between time, money, and quality for achieving the most balanced result.

7.2.2. Scientific

The scientific contribution is found when the research has an assumption that ANN is the best technique in developing predictive models. However, this research found that ANN is not the ultimate technique to develop predictive models. The complexity of a method will not always give a better result than other regression techniques. The evaluation of regression techniques shows that each technique has strength and weakness. Thus, the user needs to have a clear objective and use the regression techniques wisely (*see section 5.4. Evaluation of Regression Techniques*).

Another scientific contribution is found in data exploration compared with the traditional method of predicting haul time and return time. Figure 35, the pairs plot of the automated data, shows a regression line with a small confidence interval between variable distance and haul time and return time. Figure 91 shows the mathematic equation for each regression where y refers to the distance (km).

Equation 20. Haul Time from automated data

$$Haul\ Time\ (Sec) = 175.6y - 81.77$$

Equation 21. Return Time from automated data

$$Return\ Time\ (Sec) = 49.61y + 98.76$$

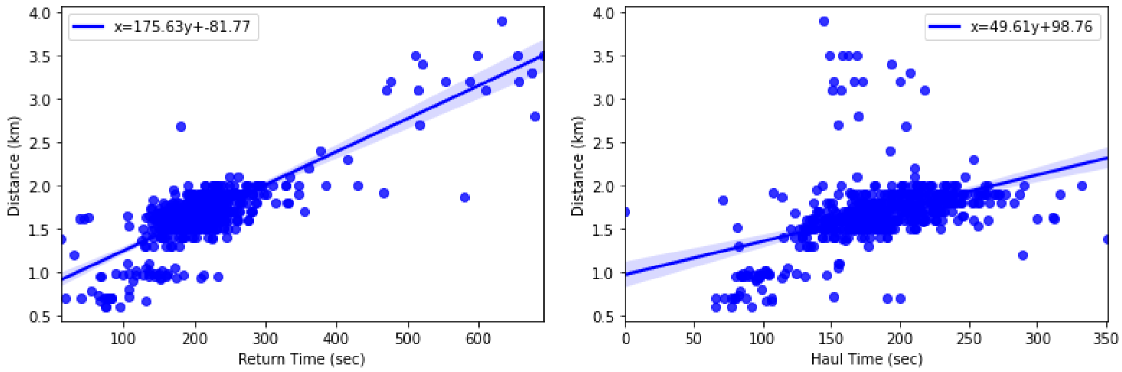


Figure 91. Regression Model for Haul Time and Return time using variable distance

Compared with Equation 2 and Equation 3, which are the traditional method in calculating haul time and return time, the traditional method can predict haul time and return time with different speed input, indicating a different type of truck model. However, the estimation of truck speed for the equation input is difficult to be obtained. Therefore, Equation 20 and Equation 21 can help predict haul time and return time where the used trucks are Caterpillar 745 and Volvo A45G.

This equation might help contractors who do not have engineers that can operate predictive models. Contractors can compare haul time and return time prediction from the equations with the output of the traditional method and improve their prediction. However, the accuracy is less than the predictive models. Thus, engineers need to check the outcome from the proposed equations.

7.3. Recommendation

This section aims to give recommendations based on the main finding and lesson learned from this research and insights from experts. Furthermore, recommendations for the application of the predictive model and further research is given as follows.

7.3.1. Practical

This research has some recommendation for experts in the construction industry, as follows.

- Raise awareness about data

Based on the discussion with experts, engineers mindset is important as the starting point for improving the construction industry. This research showed that historical data is the backbone in developing and resulting in a good predictive model. It is required an awareness of the importance of collecting and storing data in a digital platform. This awareness can lead to many opportunities in utilizing digital platforms to improve the construction industry performance

- Hire a data analyst

Data analysts with construction knowledge are recommended to be hired to maintain and develop a predictive model. Construction knowledge is important to understand the problem and data. In addition, the data analyst can develop a new predictive model for solving different problems in the construction industry.

- Improve the predictive model

Figure 92 shows the procedure to improve the predictive model, which refers to this research. Collect data is a pivotal step for developing a robust predictive model. Based on this research, the contractor lacks in documenting past projects. The data can be collected using various methods. The operation condition data can be gathered using weather API, GPS, and project documentation from different time and projects. The data about machinery condition can be collected using sensors. The sensor can help engineers to record difficult data, for instance, the volume. The data about the operation practice can be collected through KPI, where the driver's performance is recorded.

Then, the target needs to be identified to narrow down the possible method to obtain the prediction modelling, such as classification or clustering. Next, data preparation is conducted where the quality of data and the understanding of data should be done in this step. It is also important to treat the outlier carefully. Hence, analyzing the data is important because it might give a new finding.

In the modelling process, the method should be treated as a tool, not as a target. The predictive model is suggested to be developed from a simple method to a complex method. The predictive model should be tested after the development process. This procedure can help to build a robust predictive model.

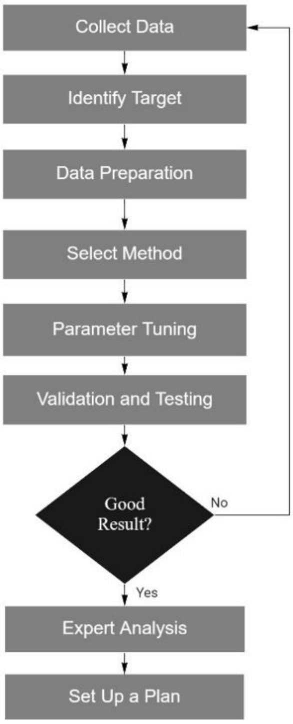


Figure 92. Practical Recommendation

7.3.2. Scientific

This research can be a starting idea for future research in improving TCT estimation in earthworks. The research recommendation can be given for future improvement, as follows.

- Explore hyperparameter tuning

This research can be developed deeper by exploring different hyperparameter, which might obtain a different result. Especially to explore different hyperparameter tuning for developing a load time model.

- Investigate the outliers

Detecting outliers also needs to be explored to understand the process in the truck cycle. This research used the statistic method for capturing the outliers. However, there might be some hidden outliers that have not yet be detected.

- Investigate the usage of synthetic data

Utilizing synthetic data can be insightful to know the minimum data size to improve the predictive model. The quantity of data needed and the type of data needed might give a vast contribution.

- Explore different methods

Research about the efficient method to attain good quality data can be explored, such as using video or camera to attain real-time data. Exploring other deep learning methods for image recognition to extract data can vastly contribute to the construction industry.

- Improve the data

The predictive model also can be improved by adding more data from different type of projects and features. Since this is not a predictive model that can learn continuously, this suggestion might achieve a more robust predictive model.

- Investigate the usage of the predictive model.

This research started with the technological approach, where machine learning is currently a popular method in many industries to improve prediction. Further research can assess the usage of this predictive model in real work. The finding might lead to the finding of engineer readiness for the application of machine learning. The proper way to collaborate can be proposed to obtain an effective improvement in the construction industry.

- Investigate different construction site

Developing predictive models of TCT in an urban area is suggested because many factors affect the prediction result, for instance, the time slot for a truck to transport the material. An accurate prediction can help contractors in ensuring the client or government about the logistics activity.

Bibliography

[1] Caterpillar, "Loading and Hauling," 2015.

[2] R. Rapp and B. Benhart, Construction Site Planning and Logistical Operations: Site Focused Management for Builders, West Lafayette, Indiana: Purdue University Press, 2015.

[3] A. H. Memon, I. A. Rahman, I. Ismail and N. Y. Zainun, "Time Management Practices in Large Construction Projects," *Colloquium on Humanities, Science and Engineering*, 2014.

[4] A. Gilchrist and A. EN, "Quantification of social costs associated with construction projects: a state-of-the-art review," *Tunnel Undergr Sp Technol*, vol. 20, no. 1, pp. 89-104, 2005.

[5] P. C. Nolz, "Optimizing construction schedules and material deliveries in city logistics: a case study from the building industry," *Flexible Service and Manufacturing Journal*, 2020.

[6] Research&Markets, "Growth Opportunities in the Global Construction Industry," 2021.

[7] R. Shah, "Earthwork Planning and Visualisation of Time-Location Information in Road Construction Projects," *Journal of Advanced College of Engineering and Management*, vol. 1, 2015.

[8] R. L. Peurifoy, C. J. Schexnayder and A. Shapira, "Trucks and Hauling Equipment," in *Construction Planning, Equipment, and Methods*, New York, Suzanne Jeans, 2006, p. 296.

[9] E. G. Cervantes, S. Upadhyay and H. Askari-Nasab, "Improvement to Production Planning in Oil Sands Mining Through Analysis and Simulation of Truck Cycle Time," Edmonton, 2018.

[10] J. Choi, J. Xuelei and W. Jeong, "Optimizing the Construction Job Site Vehicle Scheduling Problem," *Sustainability 10*, 2018.

[11] A. Ekeskar, "Exploring Third-Party Logistics and Partnering in Construction: A Supply Chain Management Perspective," Linkoping University Electronic Press, 2016.

[12] X. Xue, W. Sun and R. Liang, "A new method of a real-time dynamic forecast of truck link travel time in open mines," *Journal of the China Coal Society*, vol. 37, 2012.

[13] T. Mitchell, Machine Learning, New York: McGraw Hill, 1997.

[14] M. Huebner, W. Vach and S. l. Cessie, "A systematic approach to initial data analysis is good research practice," *The Journal of Thoracic and Cardiovascular Surgery*, vol. 151, 2016.

[15] R. Plaistowe and M. Algeo, "The determination of haulage-truck requirement for an open-pit operation," *Journal of the South African Institute of Mining and Metallurgy*, 1979.

[16] A. Curi, W. S. Felsch, E. d. C. Rodovlho and B. P. Meireles, "Evaluation of Haul Trucks Performance in a CSN Mine," in *MPES Conference*, 2014.

[17] X. Sun, H. Zhang, F. Tian and L. Yang, "The use of machine learning method to predict the real-time link travel time of Open-Pit Trucks," *Mathematical Problem in Engineering,* 2018.

[18] C. N. Burt, "An optimisation approach to materials handling in surface mines," Curtin University of Technology, 2008.

[19] ILO, Guide to the Training of Supervisors-Trainees' Manual/Part 1, 1981.

[20] K. Awuah-Offei, H. A. Nasab and B. Osei, "Modeling truck/Shovel energy efficiency under uncertainty," *Transaction of the Society for Mining, Metallurgy, and Exploration,* vol. 330, pp. 573-584, 2011.

[21] A. Ng, "Coursera," Coursera, [Online]. Available: https://www.coursera.org/learn/deep-neural-network/lecture/cxG1s/train-dev-test-sets. [Accessed 10 April 2021].

[22] A. Bhande, "Medium," Grey Atom, 11 March 2018. [Online]. Available: https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76. [Accessed 19 April 2021].

[23] S. Haykin, Neural Networks and Learning Machines, Pearson, 3rd edition, 2011.

[24] A. Aklson, "Introduction to Deep Learning with Keras," IBM, [Online]. Available: https://www.coursera.org/learn/introduction-to-deep-learning-with-keras/lecture/HOKuA/artificial-neural-networks. [Accessed 2 February 2021].

[25] I. Dabbura, "Toward Data Science," 1 April 2018. [Online]. Available: https://towardsdatascience.com/coding-neural-network-forward-propagation-and-backpropagtion-ccf8cf369f76. [Accessed 21 April 2021].

[26] D. P. Kingma and J. L. Ba, "Adam: a Method for Stochastic Optimization," in *ICLR*, 2015.

[27] G. C. Cawley and N. L. Talbot, "On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation," *Machine Learning Research,* pp. 2079-2107, 2010.

[28] DataTechNotes, 14 2 2019. [Online]. Available: https://www.datatechnotes.com/2019/02/regression-model-accuracy-mae-mse-rmse.html. [Accessed 31 1 2021].

[29] P. R. H. A. E. Cohen, "How Evaluation Guides AI Research: The Message still Counts More than the Medium," AI Magazine, 1988.

[30] C. Rudin and J. Radin, "Harvard Data Science Review," 22 November 2019. [Online]. Available: https://hdsr.mitpress.mit.edu/pub/f9kuryi8/release/6. [Accessed 22 June 2021].

[31] M. Galarnyk, "Toward Data Science," Towards Data Science Inc., 12 September 2018. [Online]. Available: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51. [Accessed 17 April 2021].

[32] Q. Sun, "Road running time statistics method in truck scheduling," *Opencast Coal Mining Technology, vol 01,* pp. 35-37, 1998.

[33] A. Soofastaei, S. Aminossadati, M. Arefi and M. Kizil, "Development of a multi-layer perceptron artificial neural network model to determine haul trucks energy consumption," *International Journal of Mining Science and Technology 26,* 2016.

[34] M. Christopher, Logistics and Supply Chain Management: Strategies for reducing cost and improving service, New York: Prentice Hall, 1998.

[35] K. Schabowicz and B. Hola, "Mathematical-neural model for assessing productivity of earthmoving machinery," *Journal of civil engineering and management 13,* pp. 47-54, 2007.

[36] K. Erarslan, "Modelling performance and retarder chart of off-highway trucks by cubic splines for cycle time estimation," *Mining Technology, vol 114,* pp. 161-166, 2013.

[37] S. Magnusson, M. Johansson, S. Frosth, Lundberg and Kristina, "Coordinating soil and rock material in urban construction- Scenario analysis of material flows and greenhouse gas emissions," *Journal of Cleaner Production*, 2019.

Appendix 1

Box Plot

This section aims to briefly explain the box plot because it is used to analyze the data and evaluate the result. Box plot is a method to illustrate the distribution of data through their quartiles. Figure 93 shows an example of a box plot and the comparison with normal distribution, consisting of median, Q1, Q3, IQR, Minimum, Maximum, and outliers.

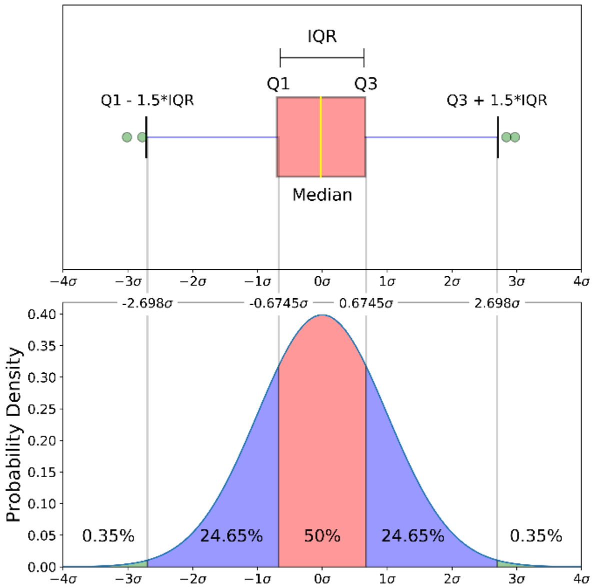


Figure 93. Box Plot [30]

The median of the 50th percentile is the middle value of a dataset. Q1 or 25th percentile is the median of the lower half of the dataset. Q3 or 75th percentile is the median of the upper half of the dataset. IQR or interquartile range is the distance between the upper and lower quartiles. The minimum or 0th percentile is the lowest data point, excluding the outliers. The maximum of 100th percentile is the largest data point, excluding the outliers. Outliers are data points that differ significantly or have abnormal distance from the other values.

Appendix 2

Hyperparameter Tuning: Manual Data

Truck Cycle Time without Return Time

Multi Linear Regression

Combination one

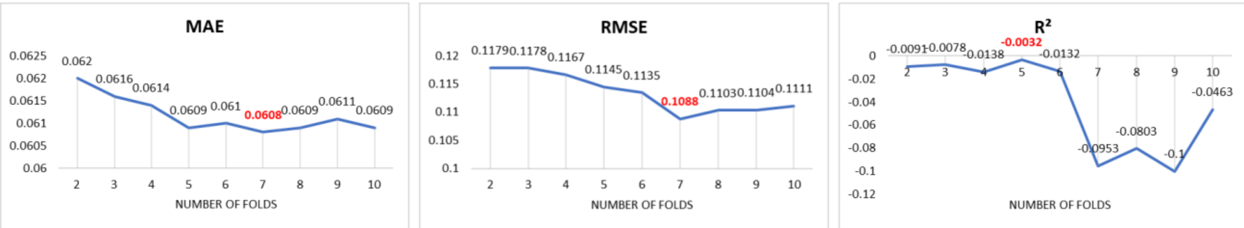


Figure 94. K-fold cross-validation for TCT model with MLR and combination one

Combination two

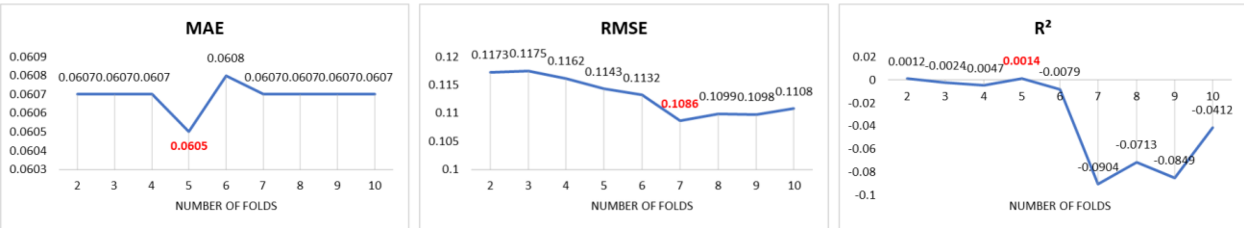


Figure 95. K-fold cross-validation for TCT model with MLR and combination two

Combination three

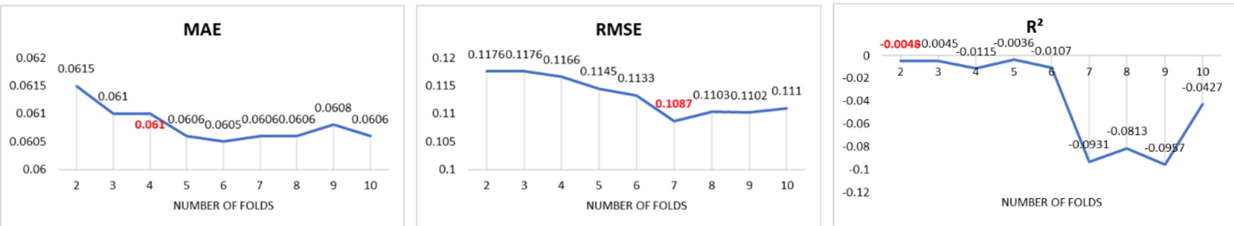


Figure 96. K-fold cross-validation for TCT model with MLR and combination three

Support Vector Regression

Combination one

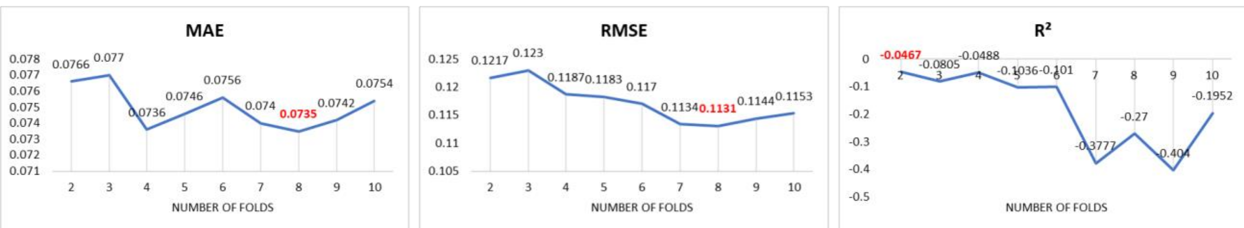


Figure 97. K-fold cross-validation for TCT model with SVR and combination one

Combination two

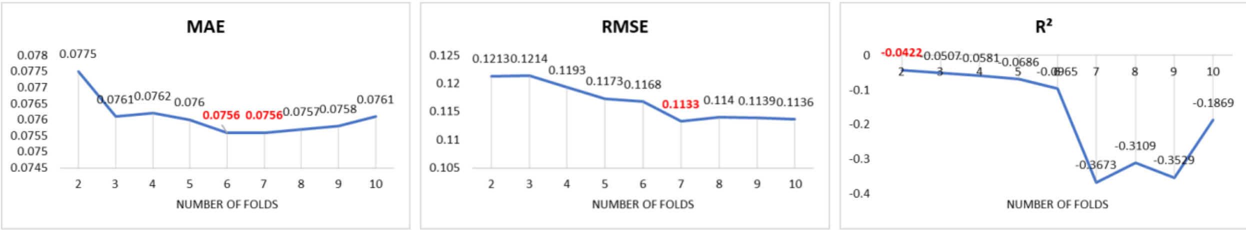


Figure 98. K-fold cross-validation for TCT model with SVR and combination two

Combination three

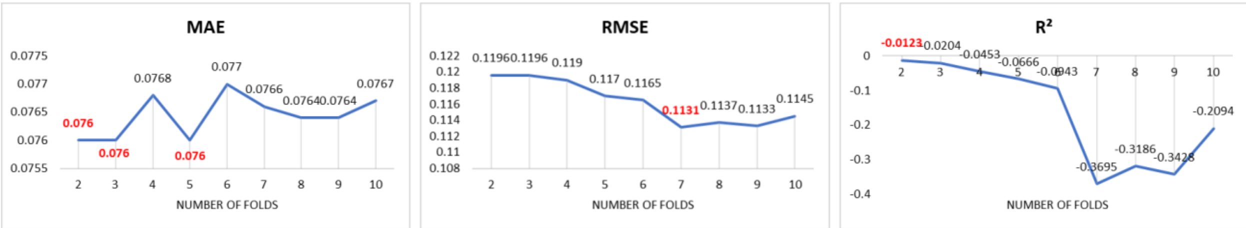


Figure 99. K-fold cross-validation for TCT model with SVR and combination three

Artificial Neural Network

Combination one

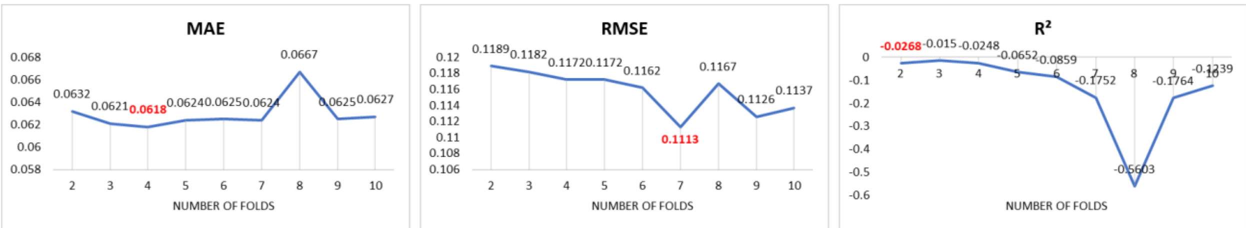


Figure 100. K-fold cross-validation for TCT model with ANN and combination one

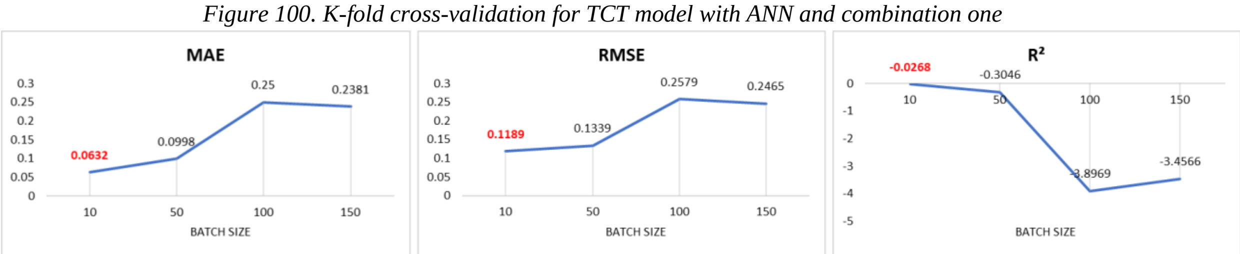


Figure 101. Batch Size for TCT model with ANN and combination one

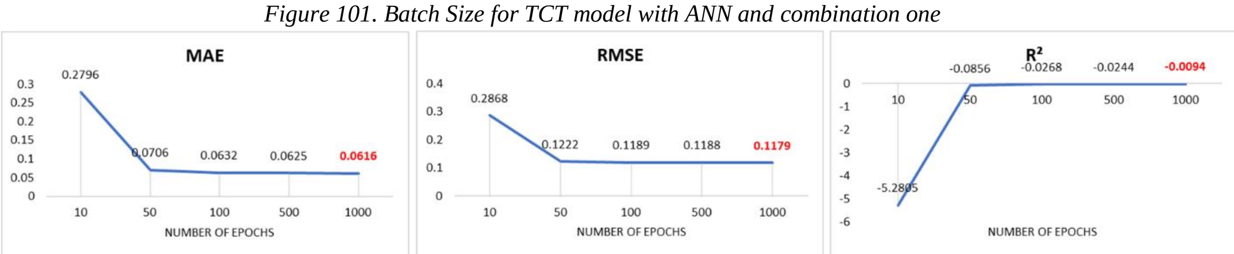


Figure 102. Epochs for TCT model with ANN and combination one

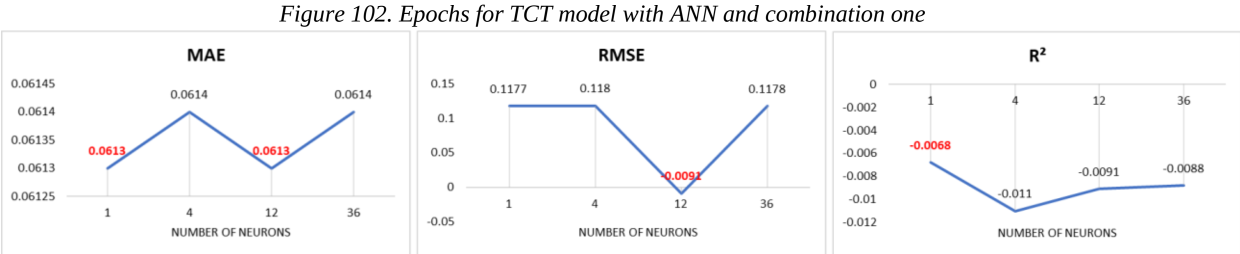


Figure 103. Neurons for TCT model with ANN and combination one

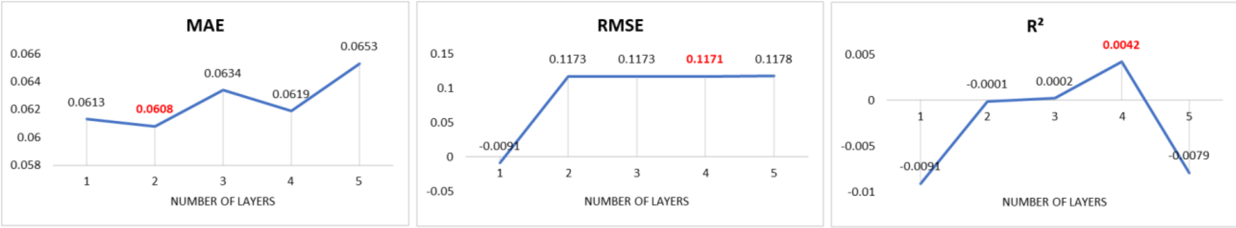


Figure 104. Hidden layers for TCT model with ANN and combination one

Combination two

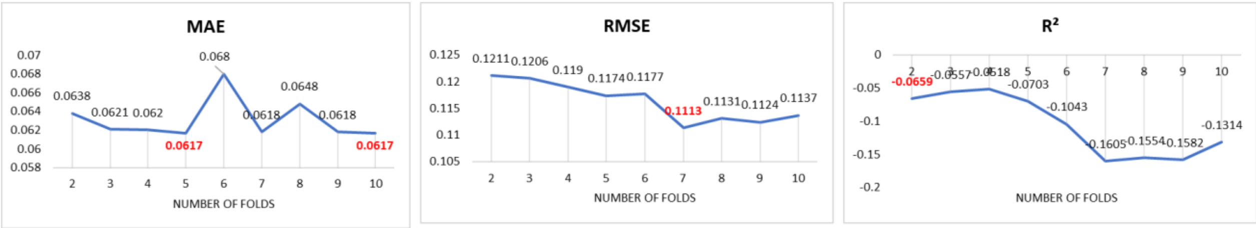


Figure 105. K-fold cross-validation for TCT model with ANN and combination two

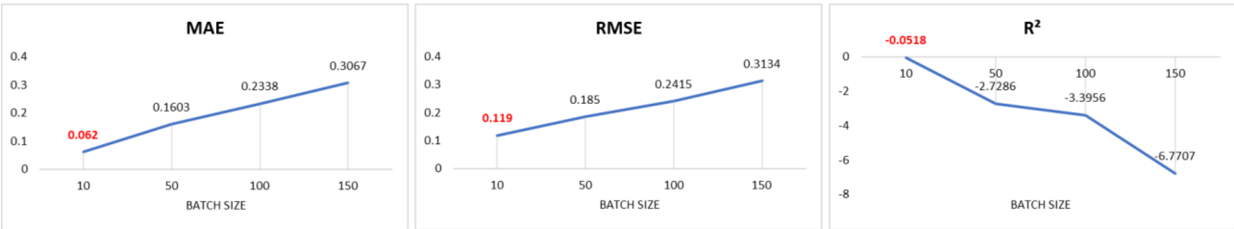


Figure 106. Batch Size for TCT model with ANN and combination two

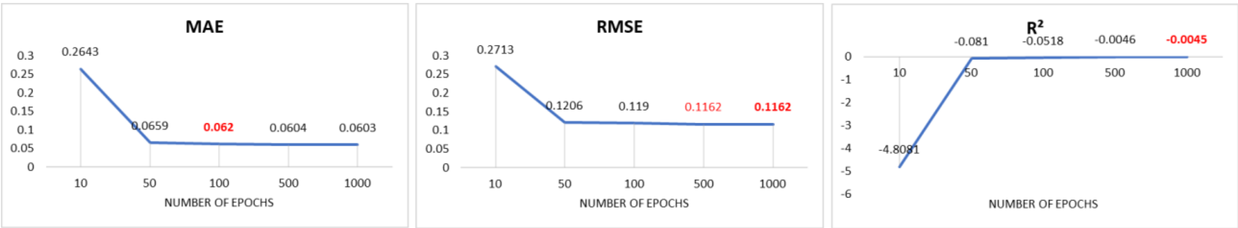


Figure 107. Epochs for TCT model with ANN and combination two

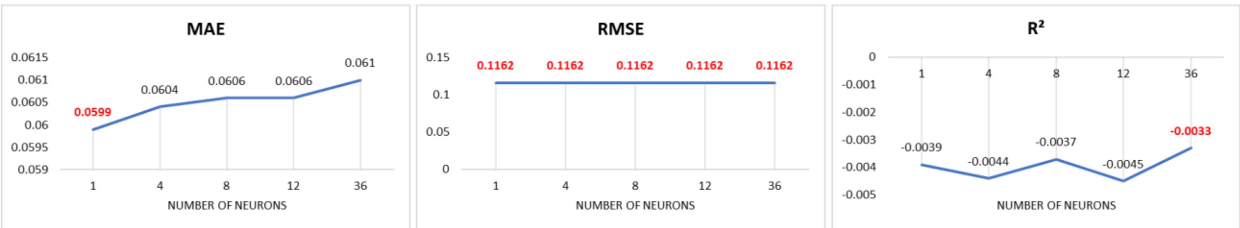


Figure 108. Neurons for TCT model with ANN and combination two

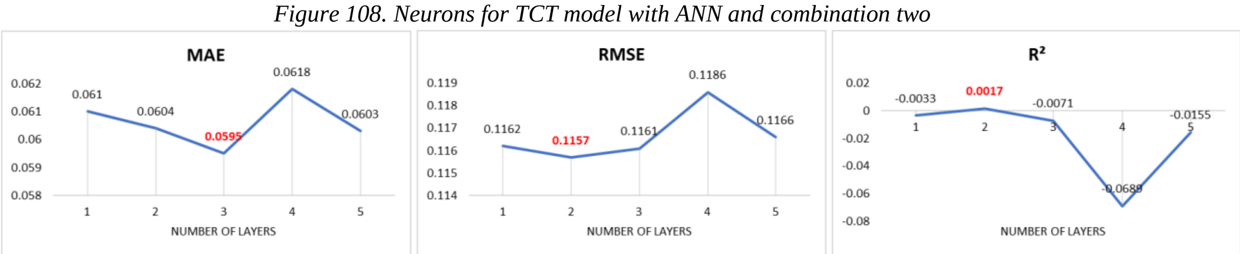


Figure 109. Hidden layers for TCT model with ANN and combination two

Combination three

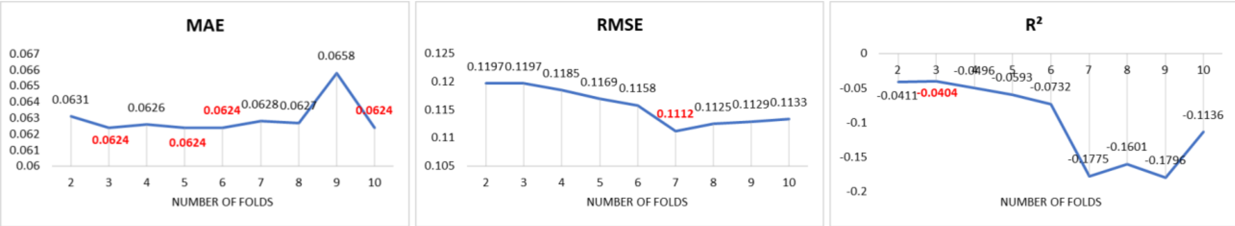


Figure 110. K-fold cross-validation for TCT model with ANN and combination three

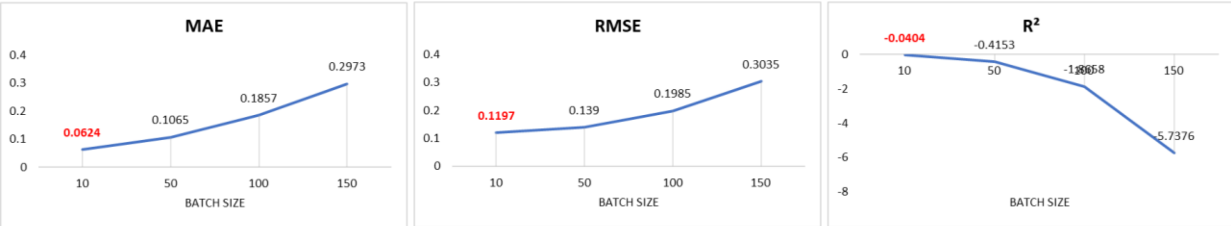


Figure 111. Batch Size for TCT model with ANN and combination three

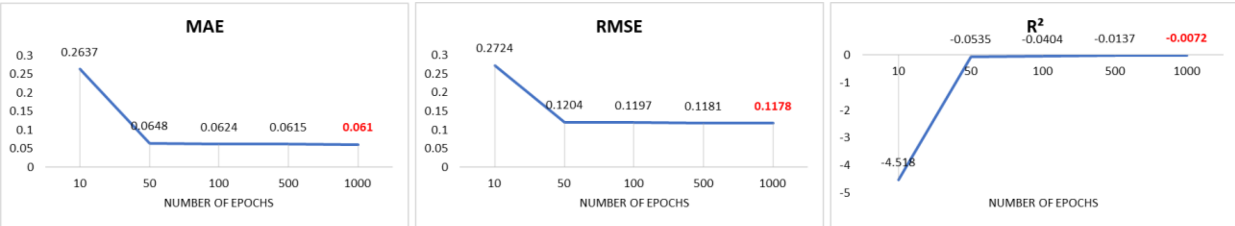


Figure 112. Epochs for TCT model with ANN and combination three

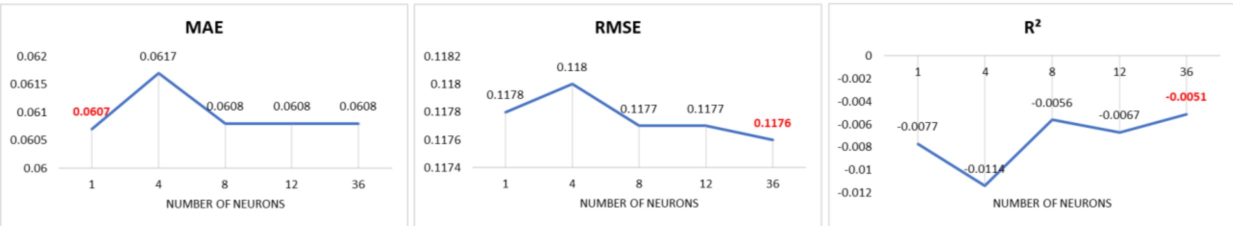


Figure 113. Neurons for TCT model with ANN and combination three

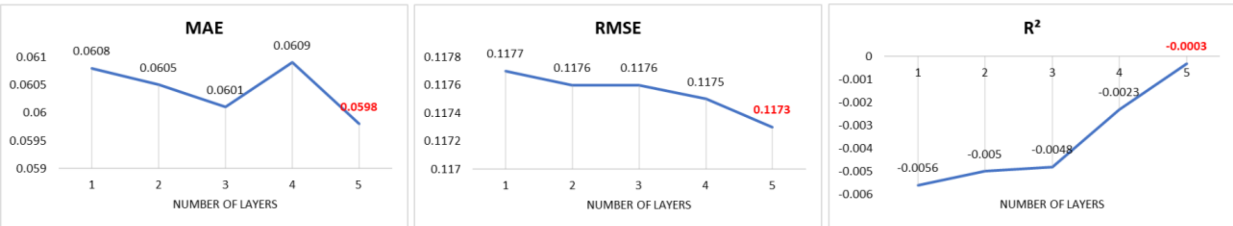


Figure 114. Hidden layers for TCT model with ANN and combination three

Hyperparameter Tuning: Automated Data

Load Time

Multi Linear Regression

Combination one

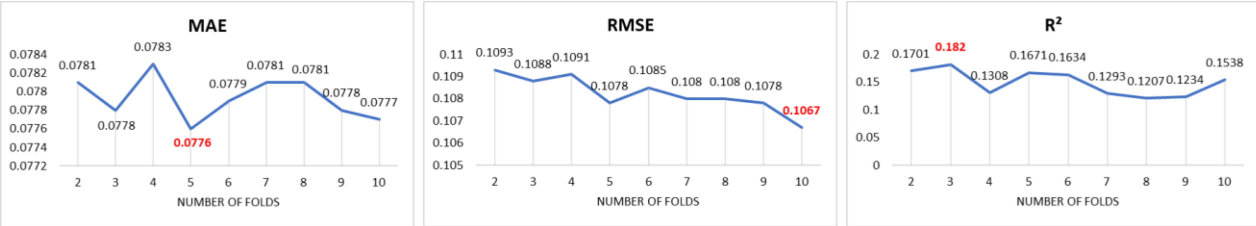


Figure 115. K-fold cross-validation for LT model with MLR and combination one

Combination two

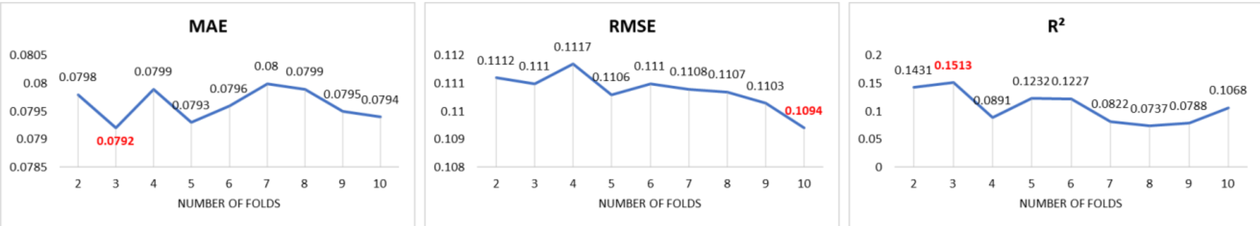


Figure 116. K-fold cross-validation for LT model with MLR and combination two

Combination two

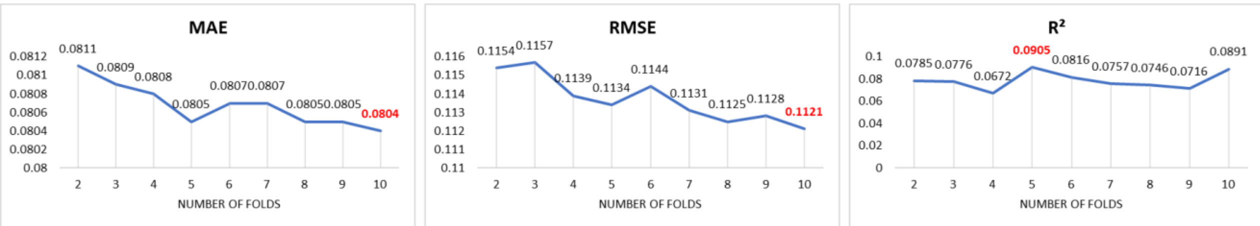


Figure 117. K-fold cross-validation for LT model with MLR and combination three

Support Vector Machine

Combination one

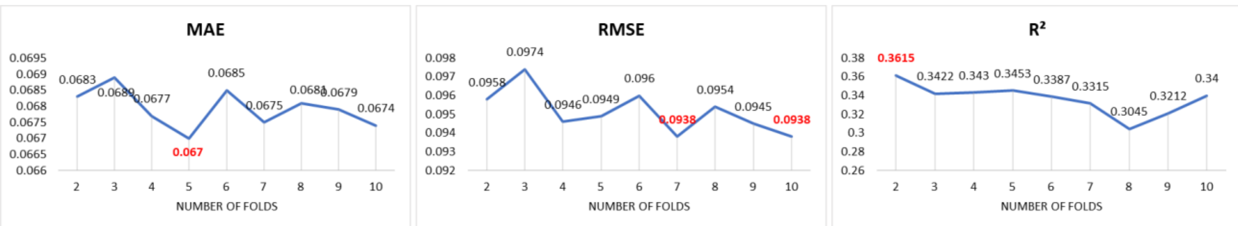


Figure 118. K-fold cross-validation for LT model with SVR and combination one

Combination two

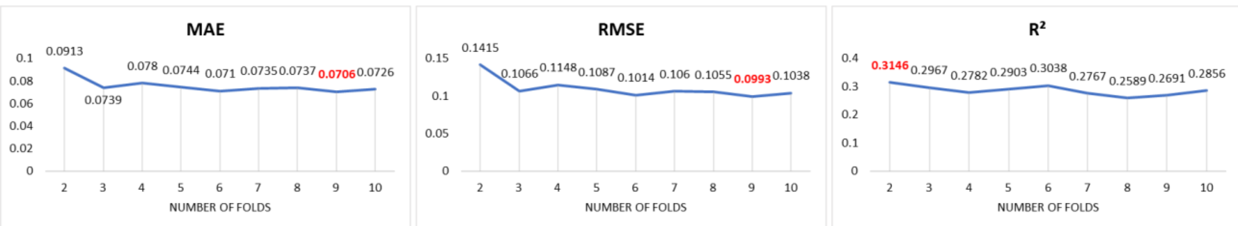


Figure 119. K-fold cross-validation for LT model with SVR and combination two

Combination three

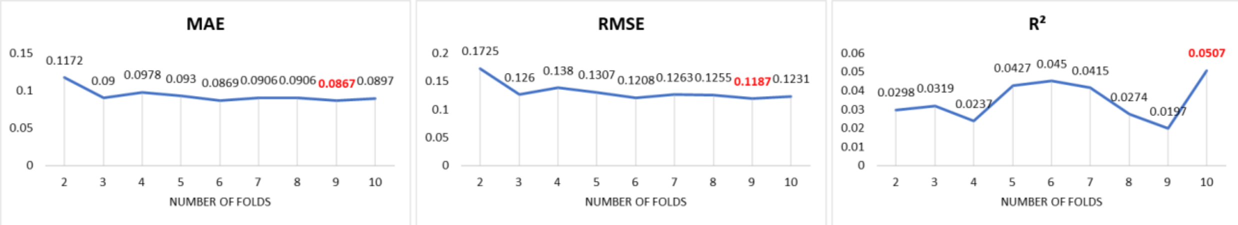


Figure 120. K-fold cross-validation for LT model with SVR and combination three

Artificial Neural Network

Combination one

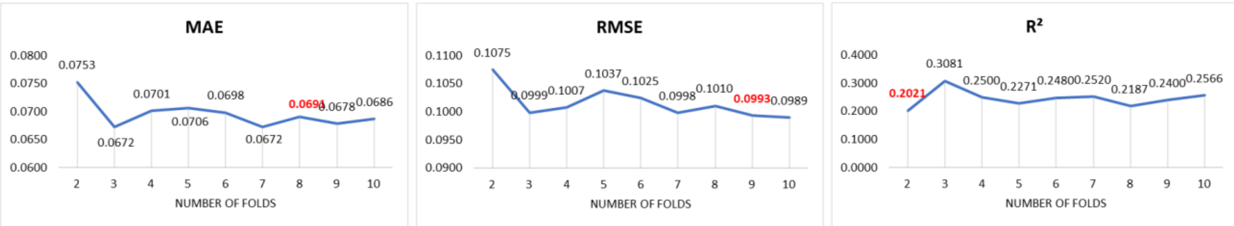


Figure 121. K-fold cross-validation for LT model with ANN and combination one

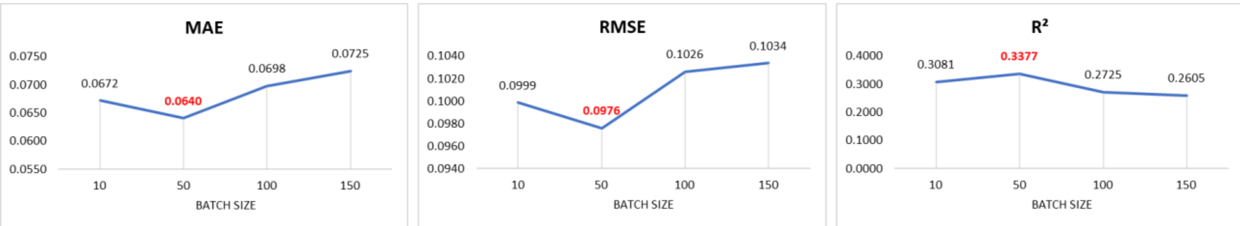


Figure 122. Batch Size for LT model with ANN and combination one

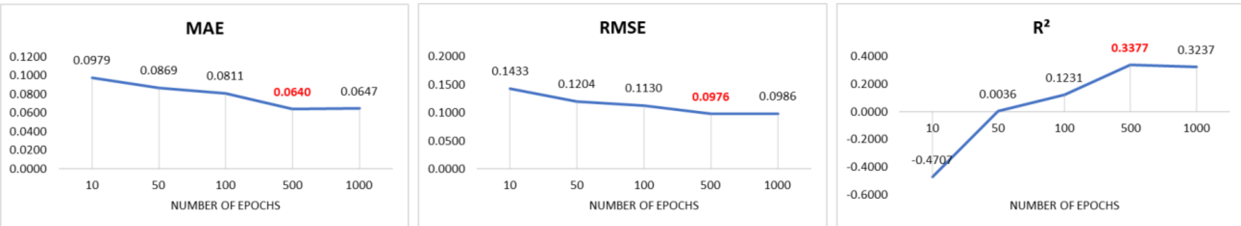


Figure 123. Epochs for LT model with ANN and combination one

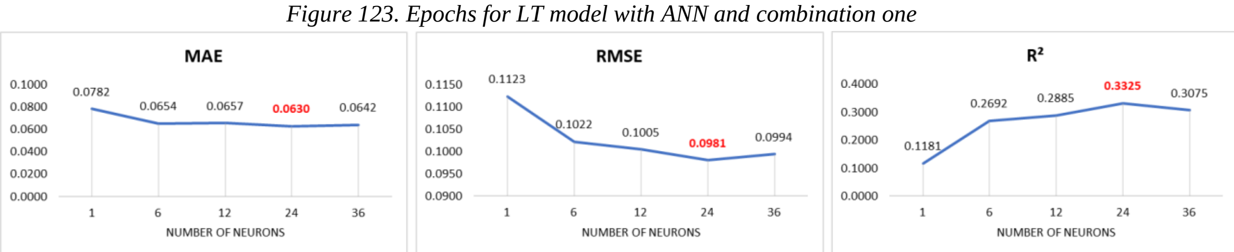


Figure 124. Neurons for LT model with ANN and combination one

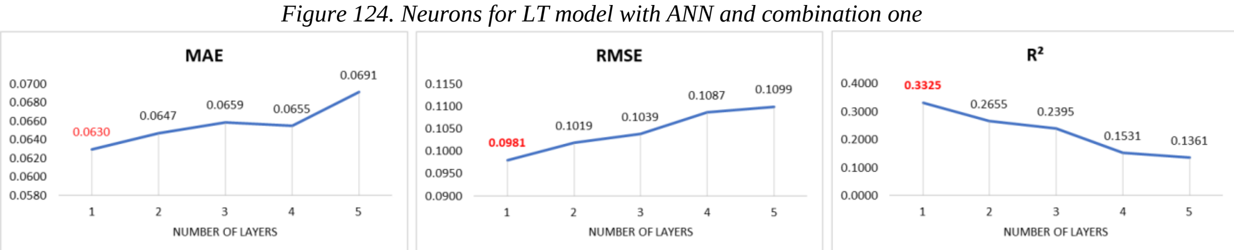


Figure 125. Hidden layers for LT model with ANN and combination one

Combination two

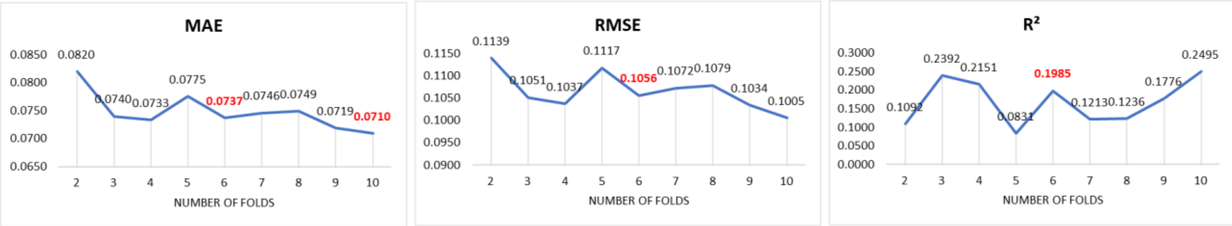


Figure 126. K-fold cross-validation for LT model with ANN and combination two

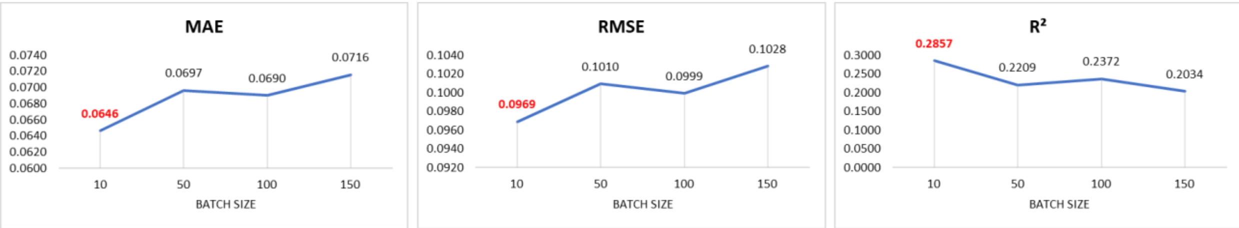


Figure 127. Batch Size for LT model with ANN and combination two

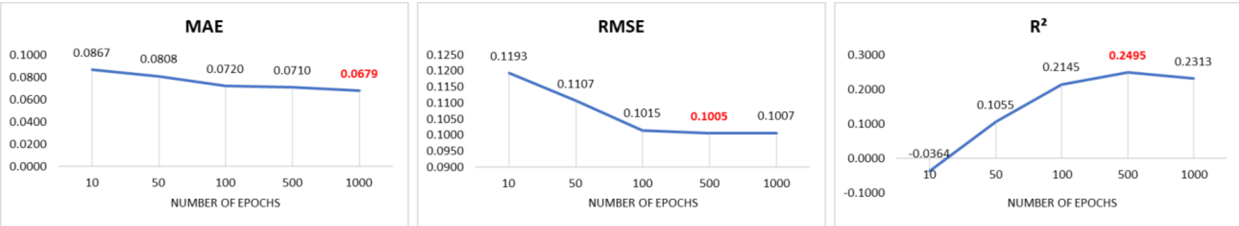


Figure 128. Epochs for LT model with ANN and combination two

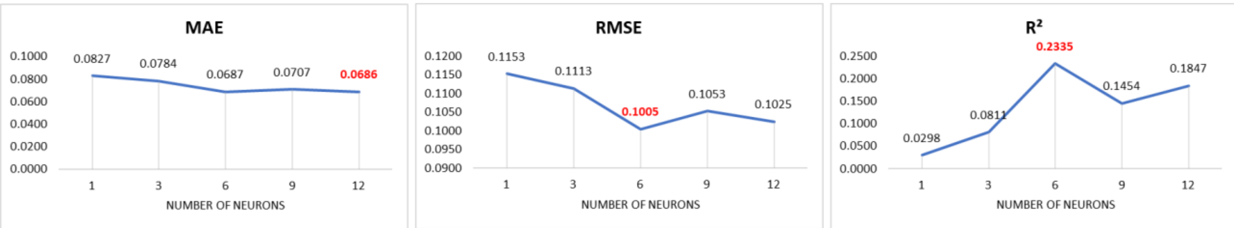


Figure 129. Neurons for LT model with ANN and combination two

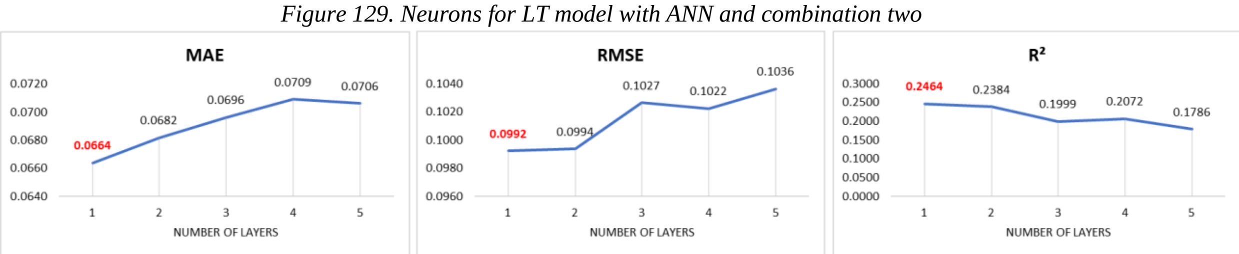


Figure 130. Hidden layers for LT model with ANN and combination two

Combination three

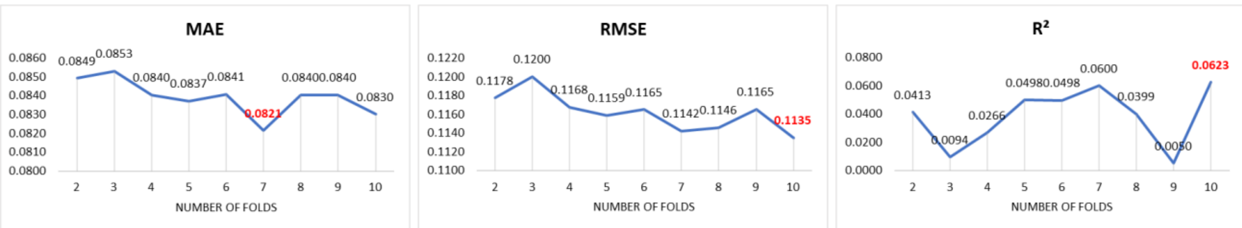


Figure 131. K-fold cross-validation for LT model with ANN and combination three

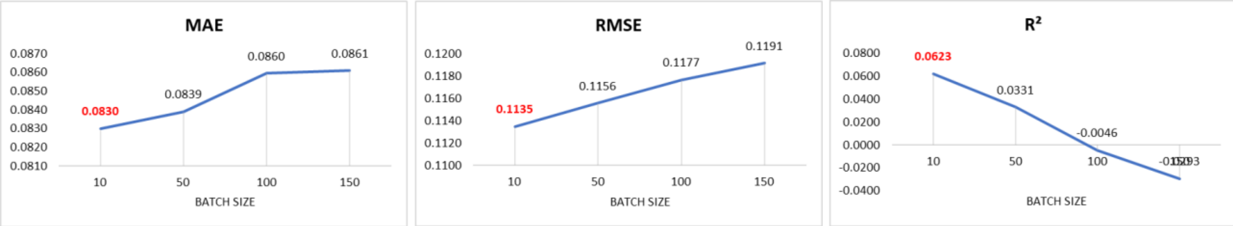


Figure 132. Batch Size for LT model with ANN and combination three

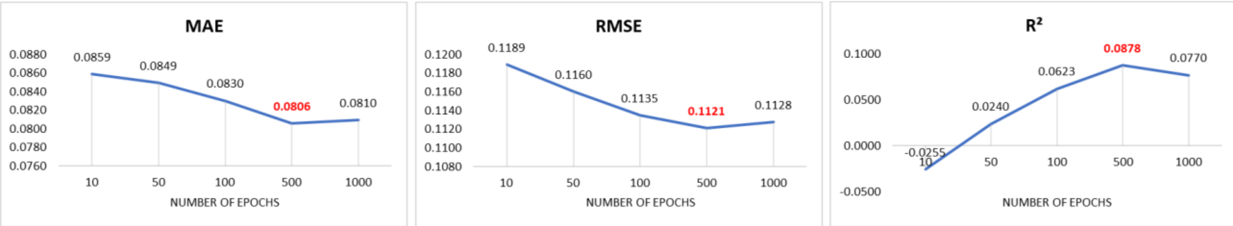


Figure 133. Epochs for LT model with ANN and combination three

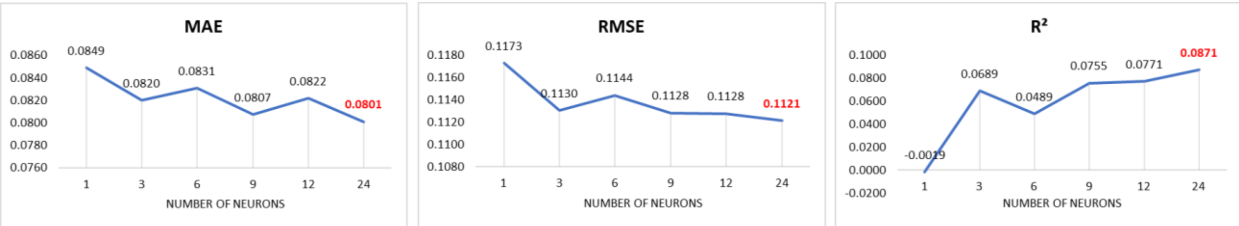


Figure 134. Neurons for LT model with ANN and combination three

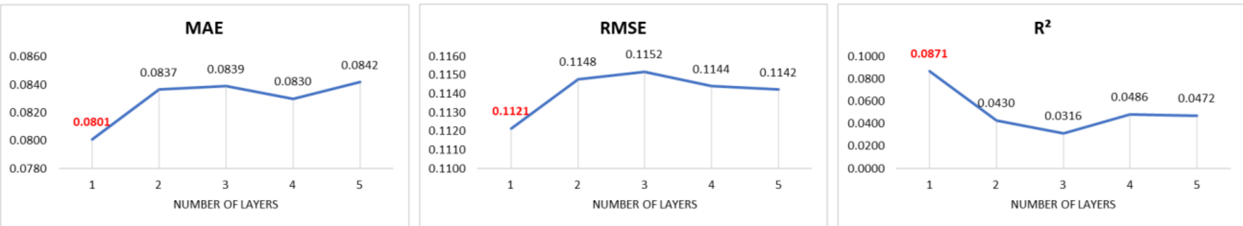


Figure 135. Hidden layers for LT model with ANN and combination three

Haul Time

Multi Linear Regression

Combination one

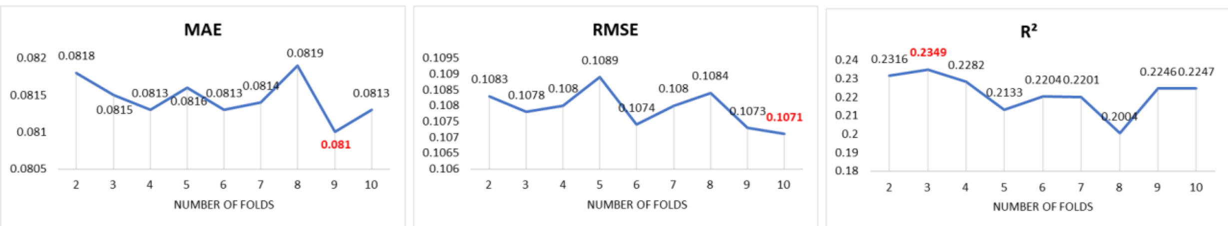


Figure 136. K-fold cross-validation for HT model with MLR and combination one

Combination two

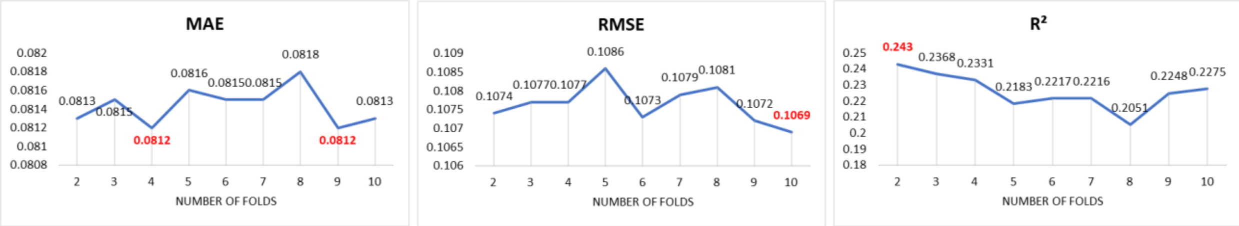


Figure 137. K-fold cross-validation for HT model with MLR and combination two

Combination three

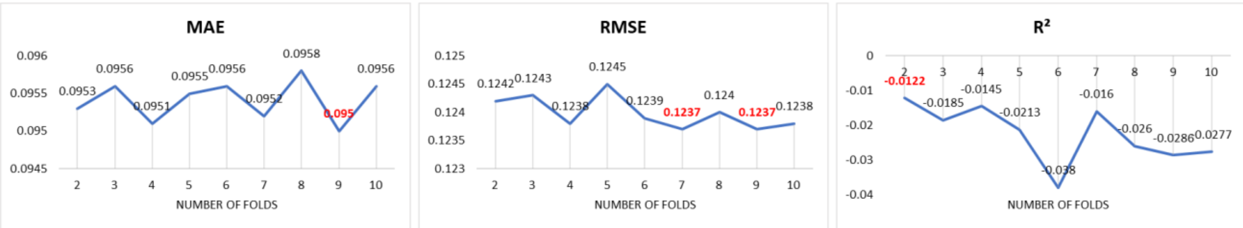


Figure 138. K-fold cross-validation for HT model with MLR and combination three

Support Vector Regression

Combination one

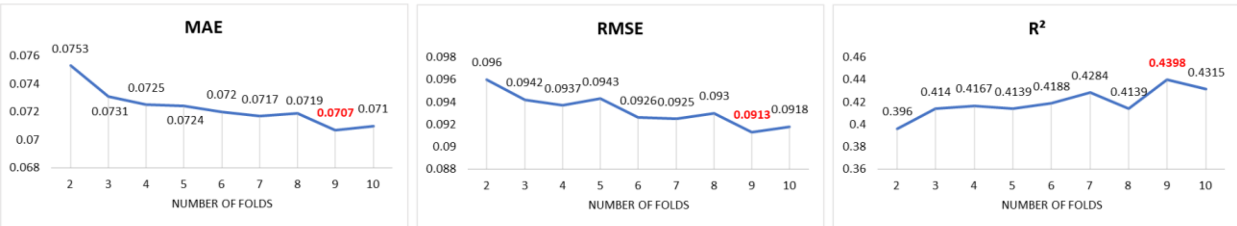


Figure 139. K-fold cross-validation for HT model with SVR and combination one

Combination two

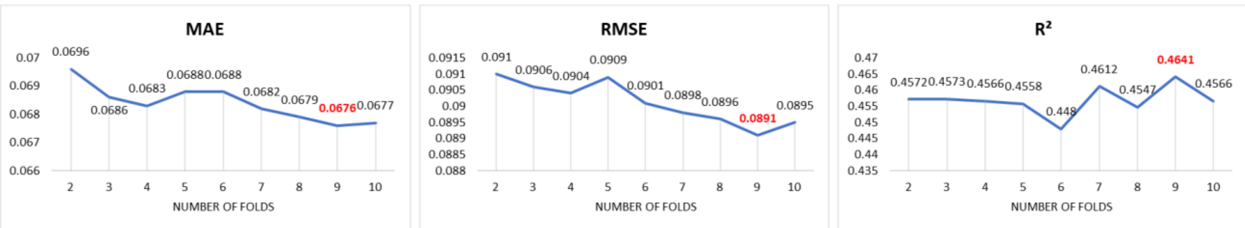


Figure 140. K-fold cross-validation for HT model with SVR and combination two

Combination three

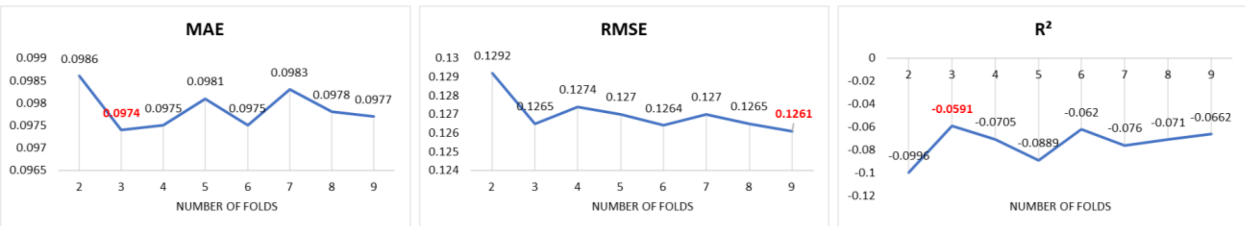


Figure 141. K-fold cross-validation for HT model with SVR and combination three

Artificial Neural Networks

Combination one

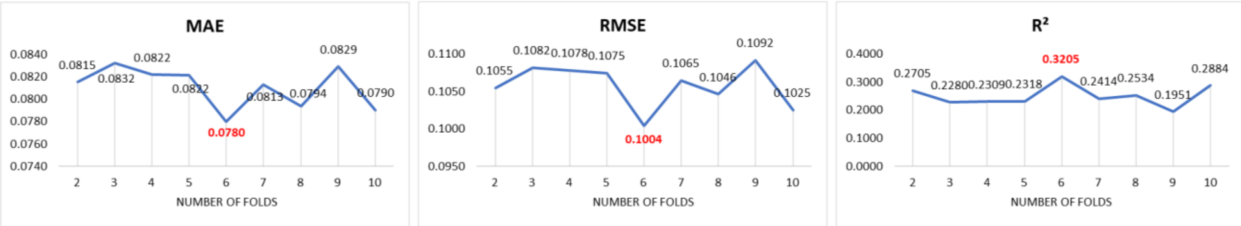


Figure 142. K-fold cross-validation for HT model with ANN and combination one

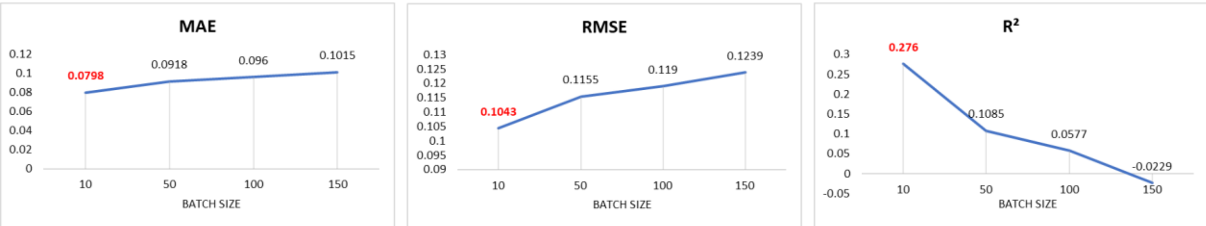


Figure 143. Batch Size for HT model with ANN and combination one

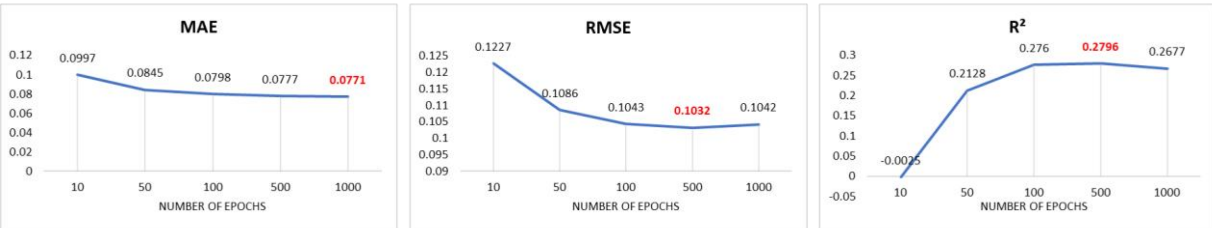


Figure 144. Epochs for HT model with ANN and combination one

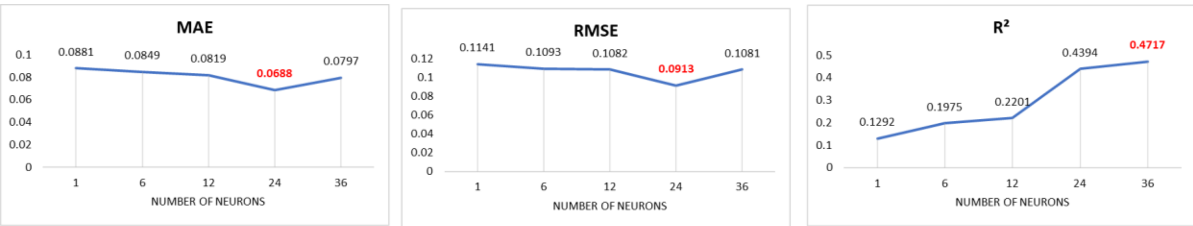


Figure 145. Neurons for HT model with ANN and combination one

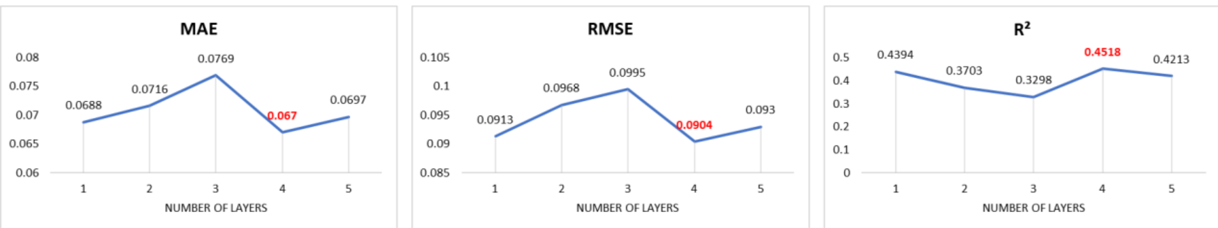


Figure 146. Hidden layers for HT model with ANN and combination one

Combination two

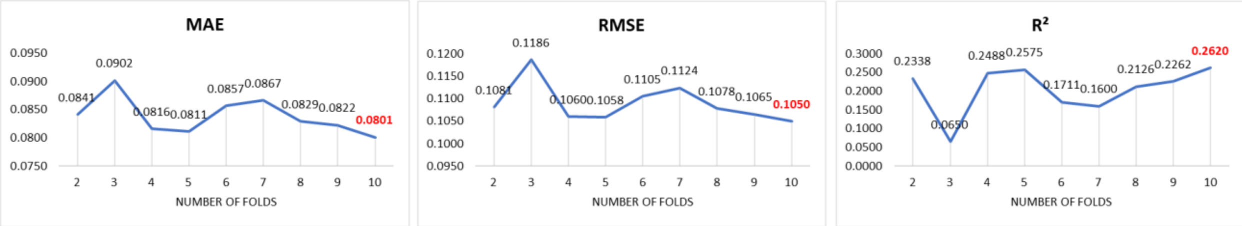


Figure 147. K-fold cross-validation for HT model with ANN and combination two

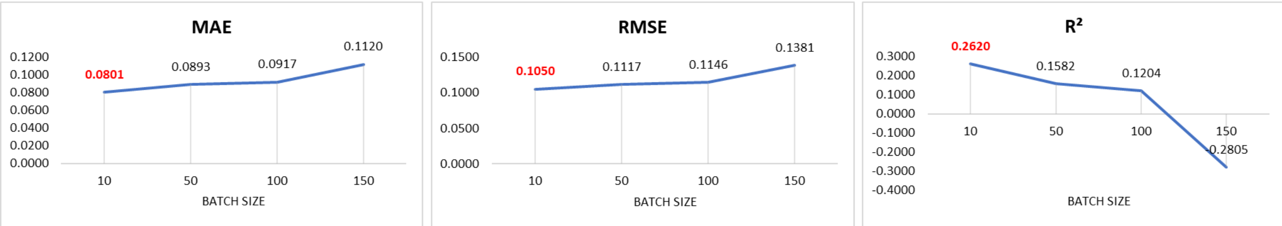


Figure 148. Batch Size for HT model with ANN and combination two

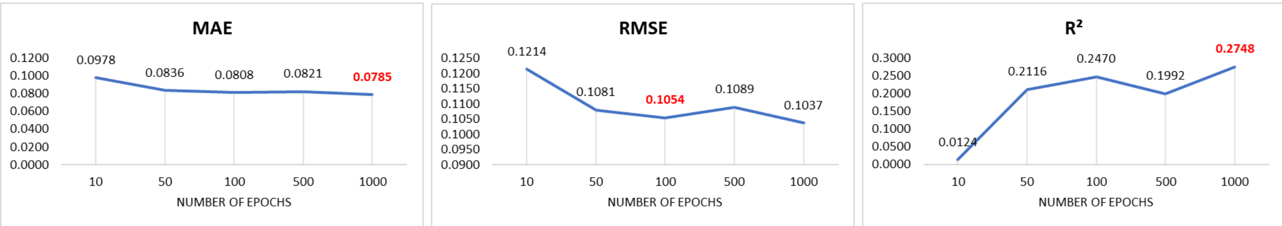


Figure 149. Epochs for HT model with ANN and combination two

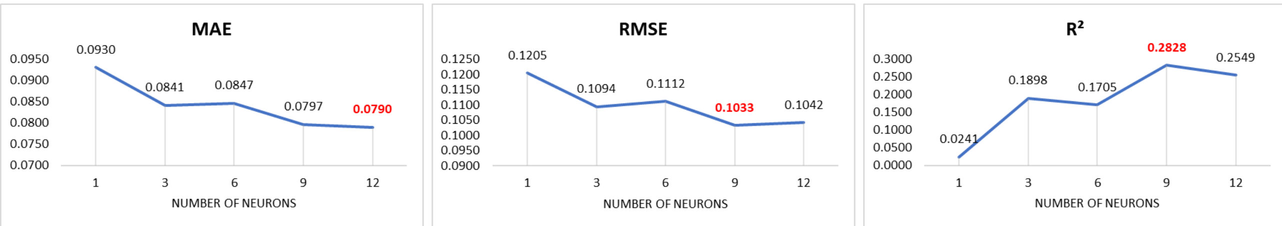


Figure 150. Neurons for HT model with ANN and combination two

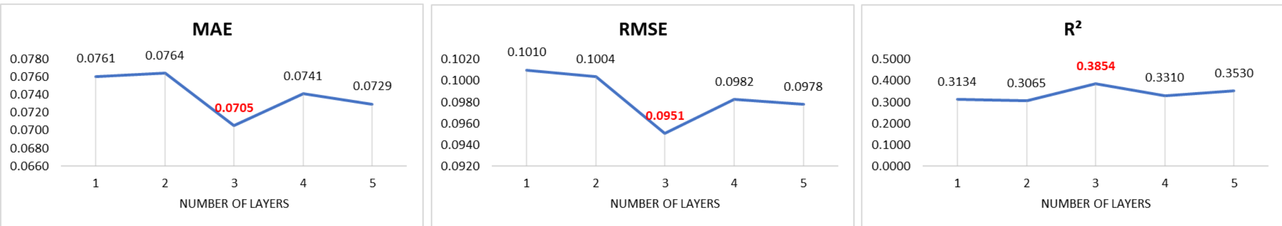


Figure 151. Hidden layers for HT model with ANN and combination two

Combination three

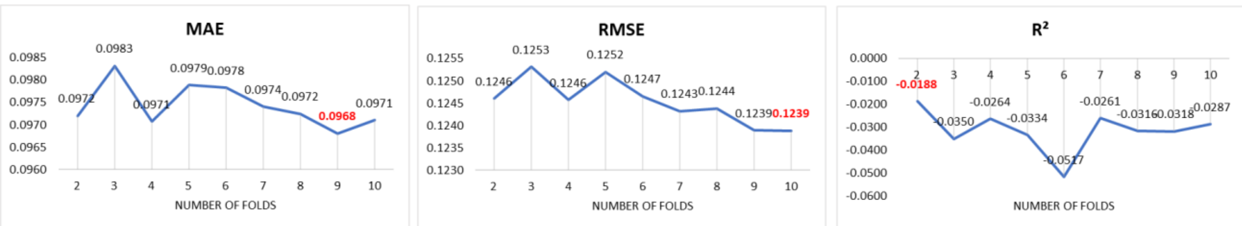


Figure 152. K-fold cross-validation for HT model with ANN and combination three

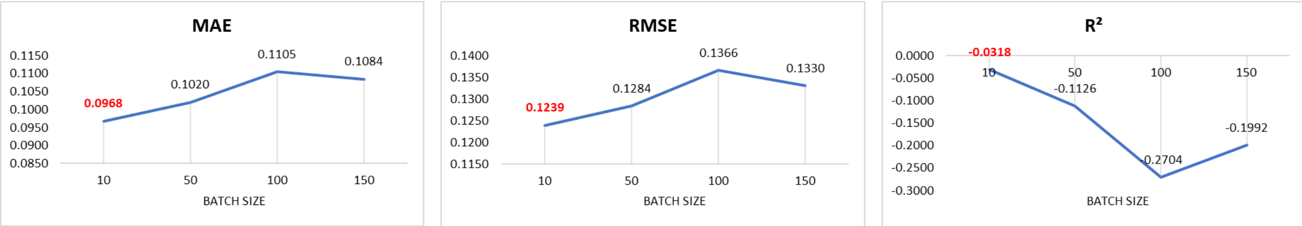


Figure 153. Batch Size for HT model with ANN and combination three

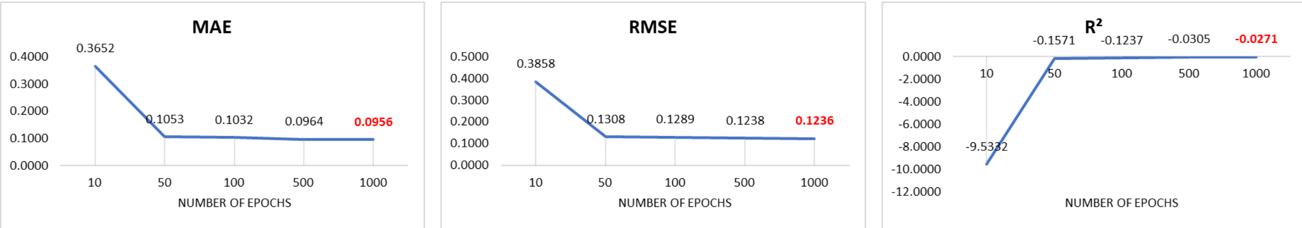


Figure 154. Epochs for HT model with ANN and combination three

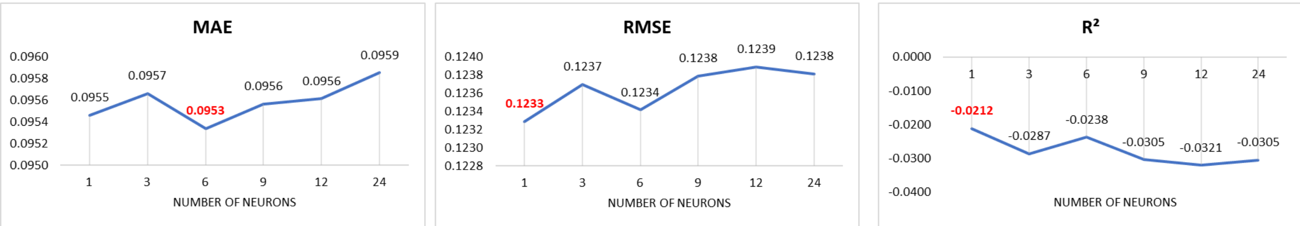


Figure 155. Neurons for HT model with ANN and combination three

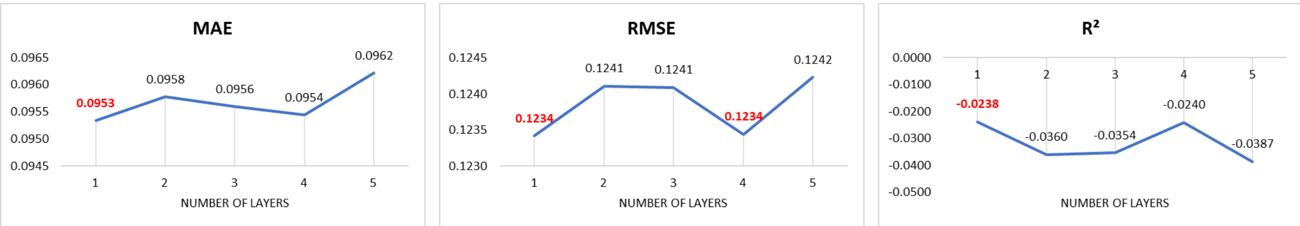


Figure 156. Hidden layers for HT model with ANN and combination three

Unload Time

Multi Linear Regression

Combination one

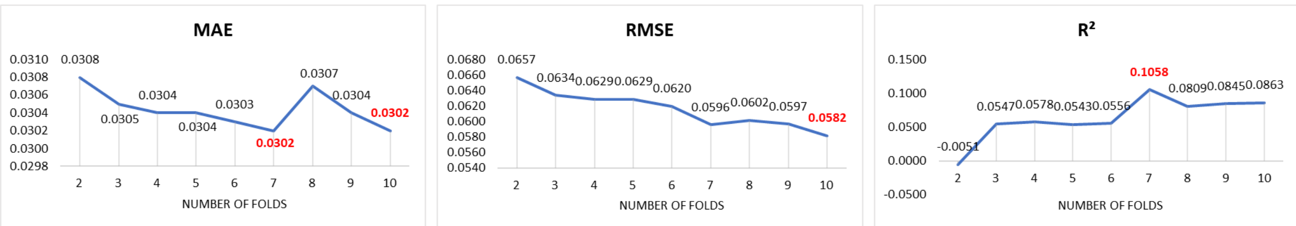


Figure 157. K-fold cross-validation for UT model with MLR and combination one

Combination two

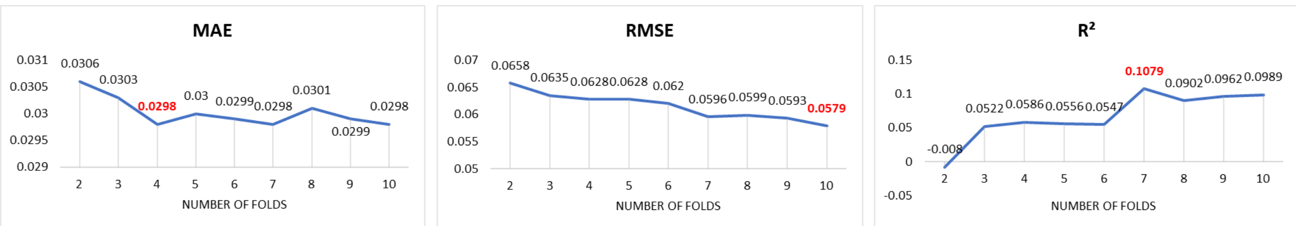


Figure 158. K-fold cross-validation for UT model with MLR and combination two

Combination three

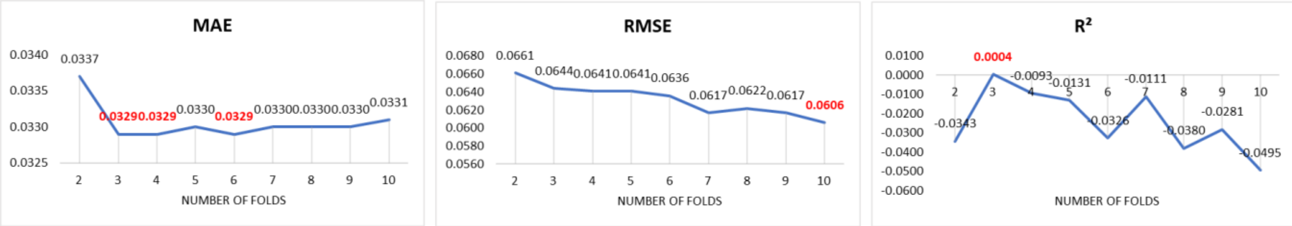


Figure 159. K-fold cross-validation for UT model with MLR and combination three

Support Vector Regression

Combination one

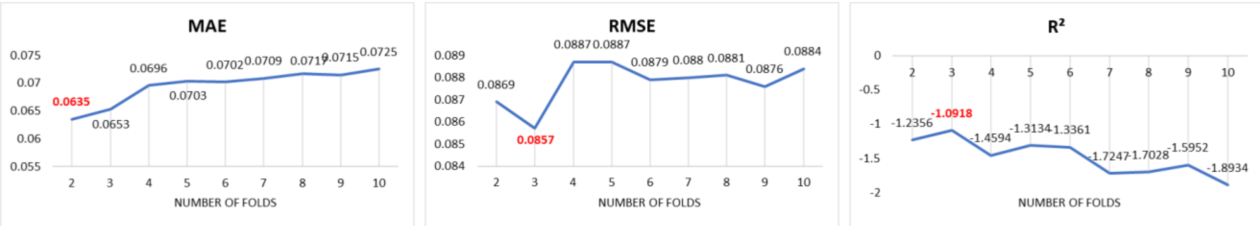


Figure 160. K-fold cross-validation for UT model with SVR and combination one

Combination two

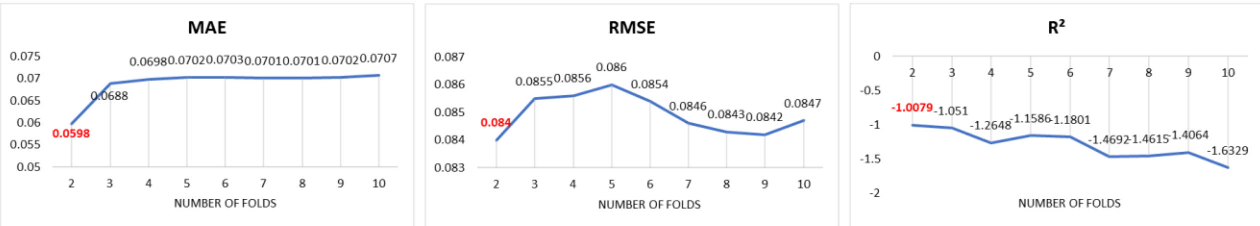


Figure 161. K-fold cross-validation for UT model with SVR and combination two

Combination three

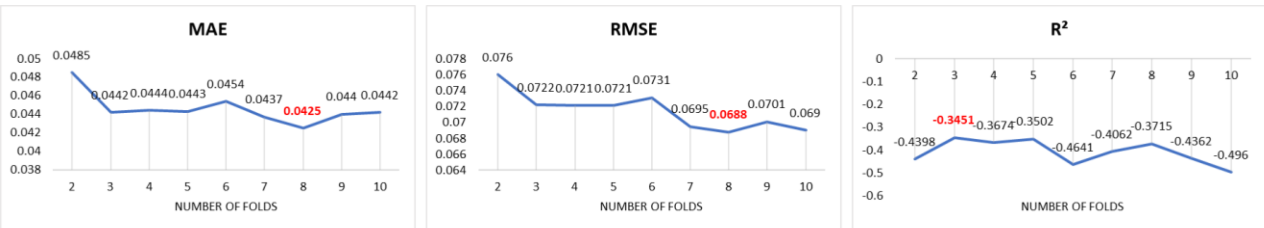


Figure 162. K-fold cross-validation for UT model with SVR and combination three

Artificial Neural Network

Combination one

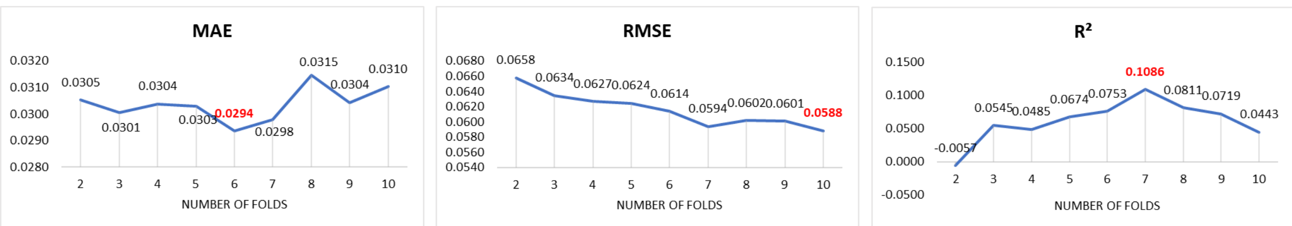


Figure 163. K-fold cross-validation for UT model with ANN and combination one

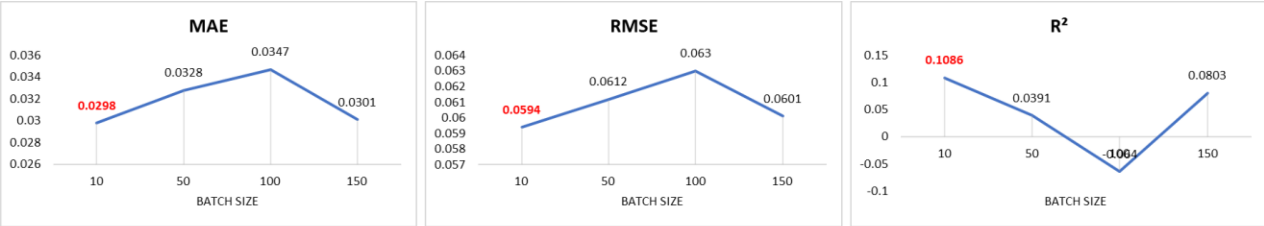


Figure 164. Batch Size for UT model with ANN and combination one

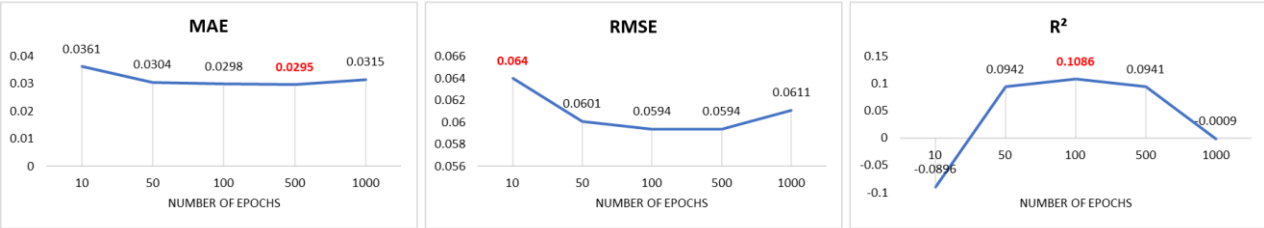


Figure 165. Epochs for UT model with ANN and combination one

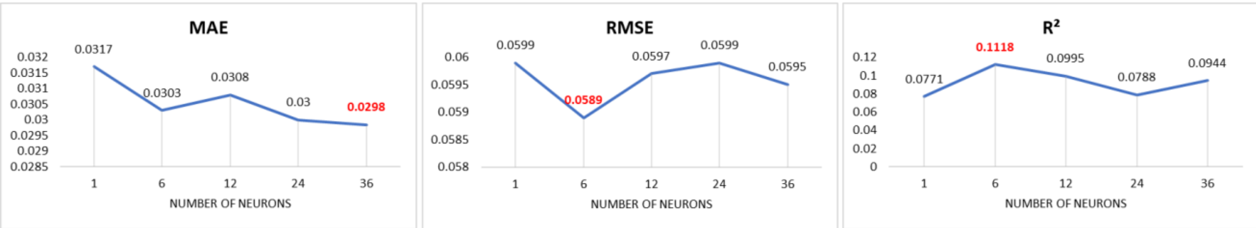


Figure 166. Neurons for UT model with ANN and combination one

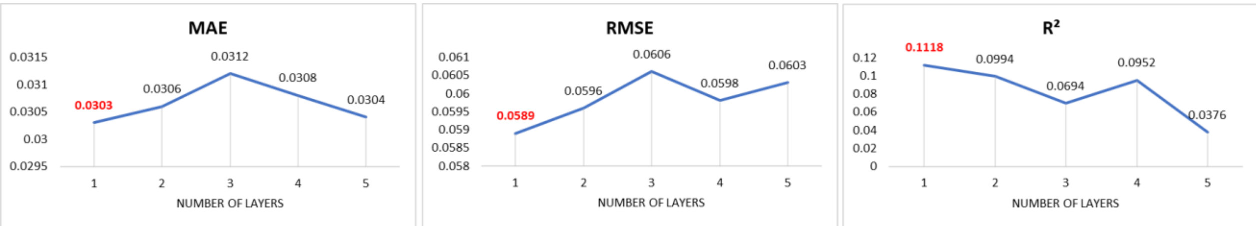


Figure 167. Hidden layers for UT model with ANN and combination one

Combination two

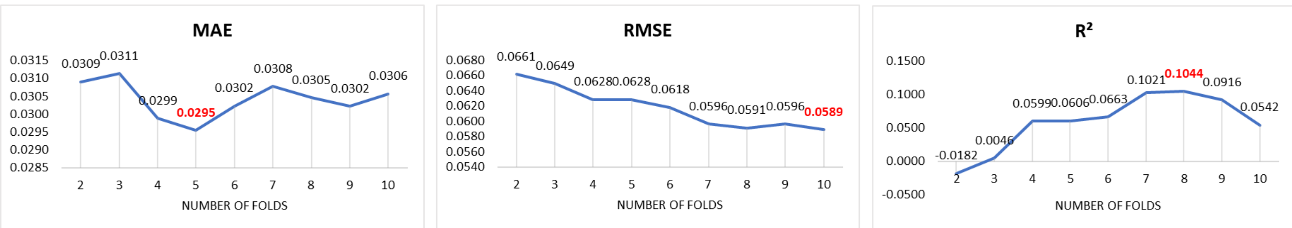


Figure 168. K-fold cross-validation for UT model with ANN and combination two

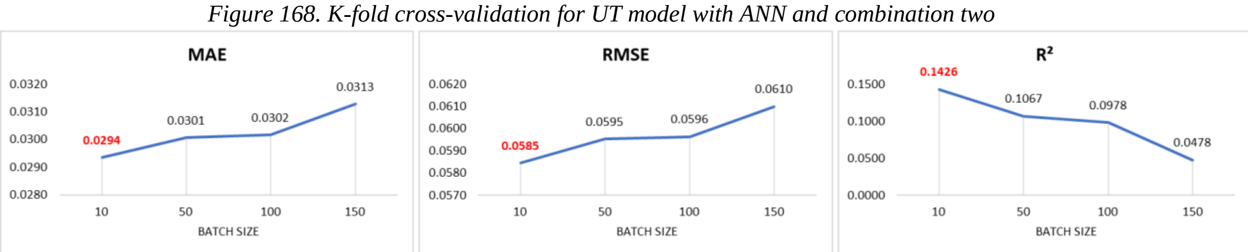


Figure 169. Batch Size for UT model with ANN and combination two

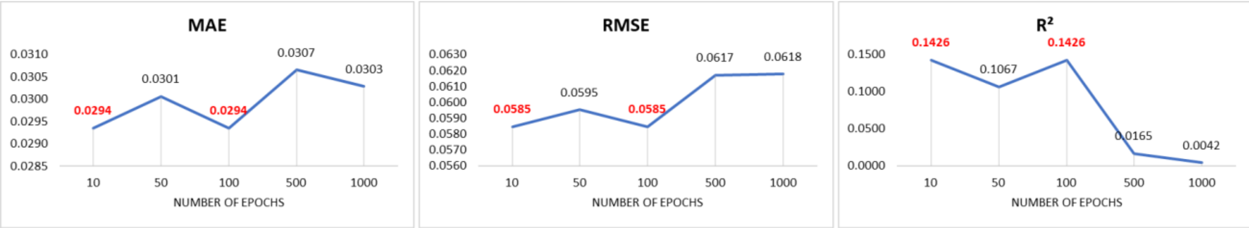


Figure 170. Epochs for UT model with ANN and combination two

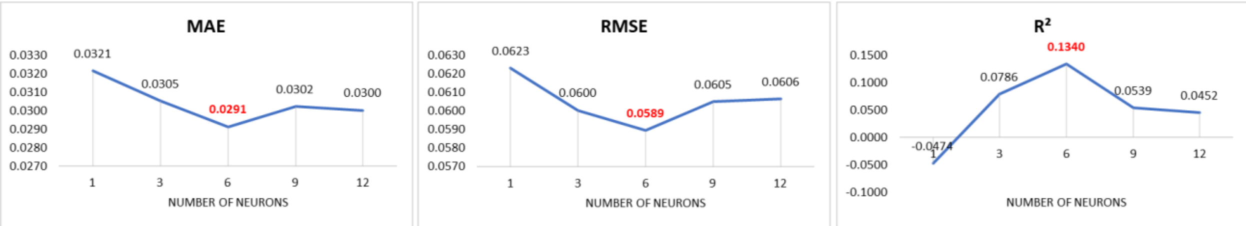


Figure 171. Neurons for UT model with ANN and combination two

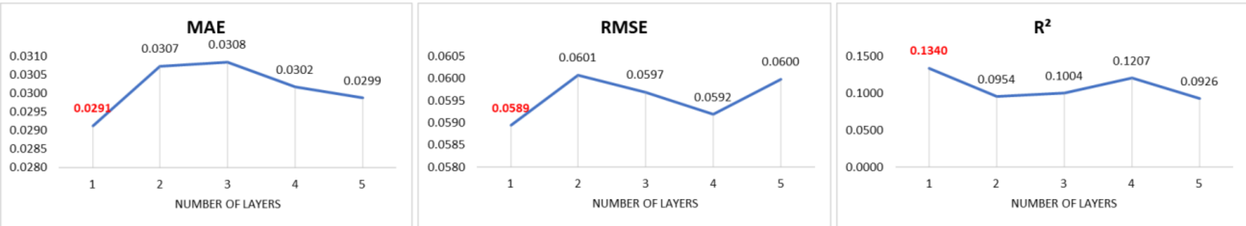


Figure 172. Hidden layers for UT model with ANN and combination two

Combination three

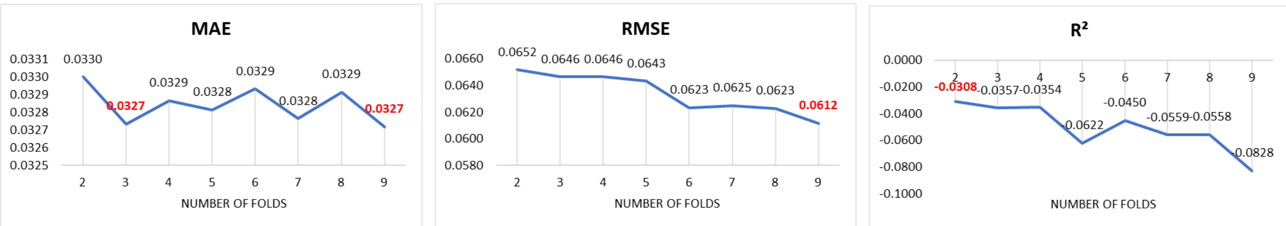


Figure 173. K-fold cross-validation for UT model with ANN and combination three

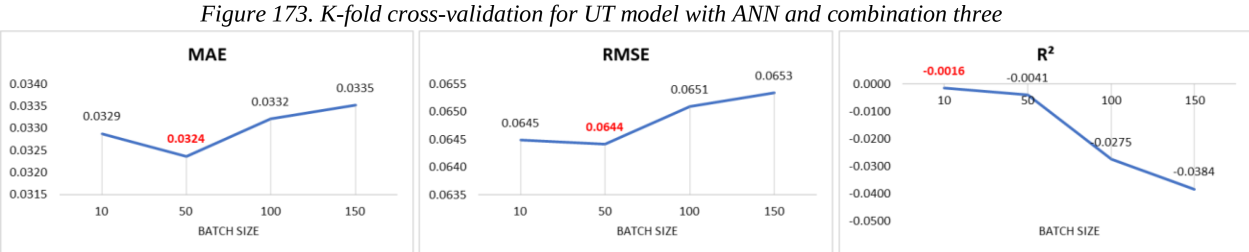


Figure 174. Batch Size for UT model with ANN and combination three

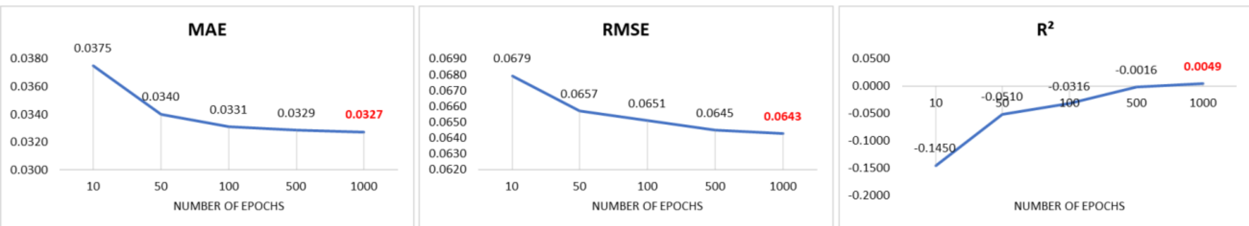


Figure 175. Epochs for UT model with ANN and combination three

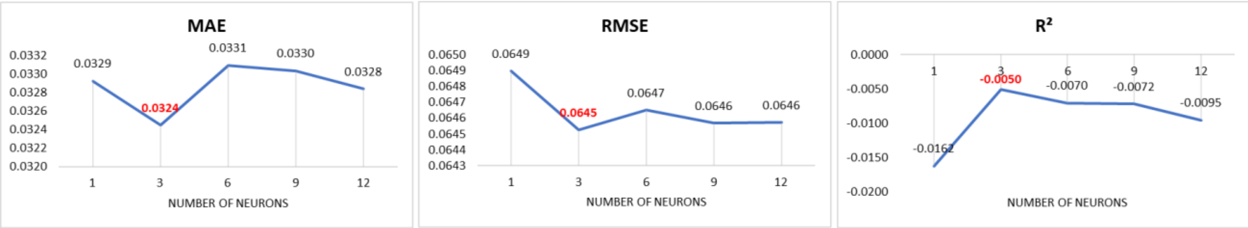


Figure 176. Neurons for UT model with ANN and combination three

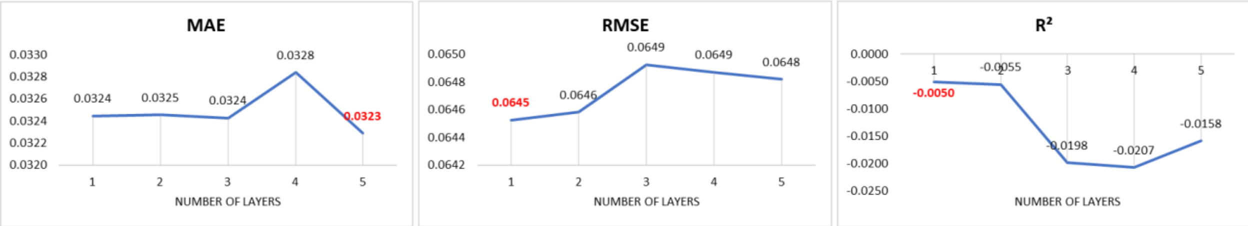


Figure 177. Hidden layers for UT model with ANN and combination three

Return Time

Multi Linear Regression

Combination one

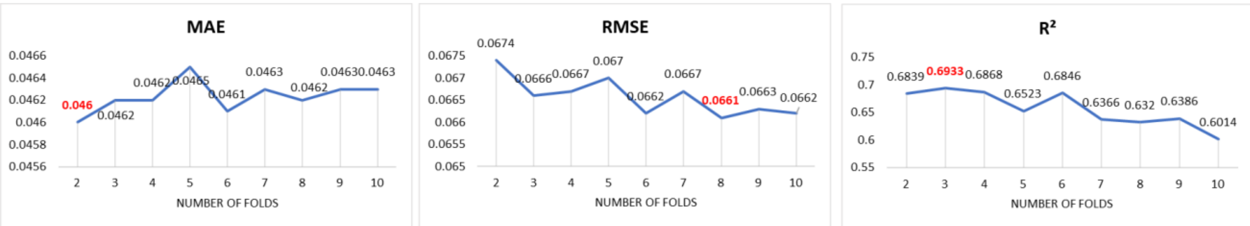


Figure 178. K-fold cross-validation for RT model with MLR and combination one

Combination two

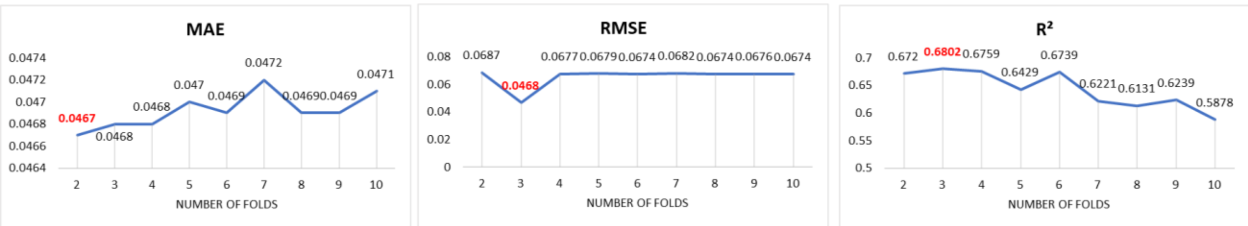


Figure 179. K-fold cross-validation for RT model with MLR and combination two

Combination three

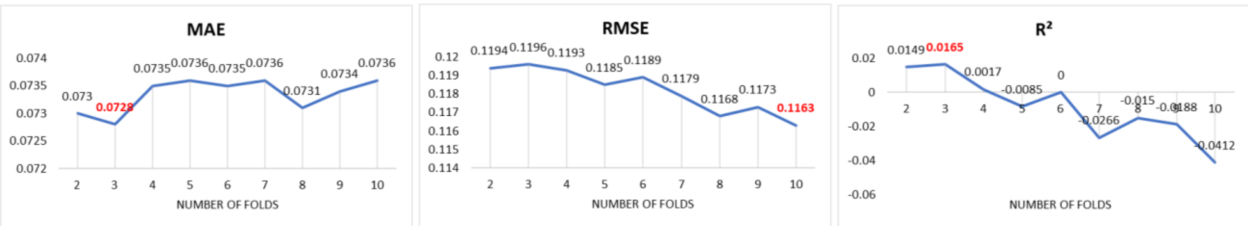


Figure 180. K-fold cross-validation for RT model with MLR and combination three

Support Vector Regression

Combination one

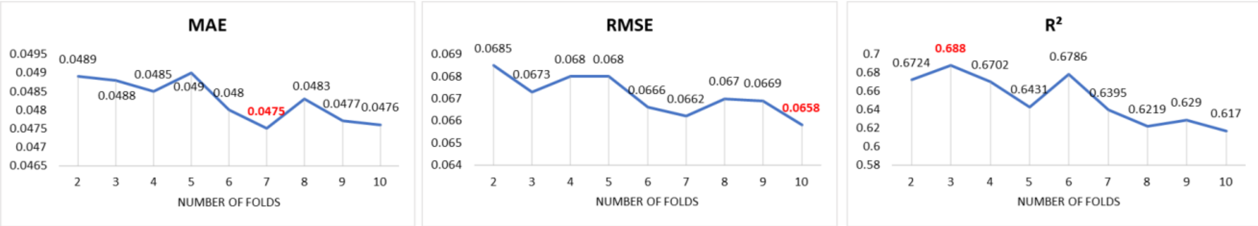


Figure 181. K-fold cross-validation for RT model with SVR and combination one

Combination two

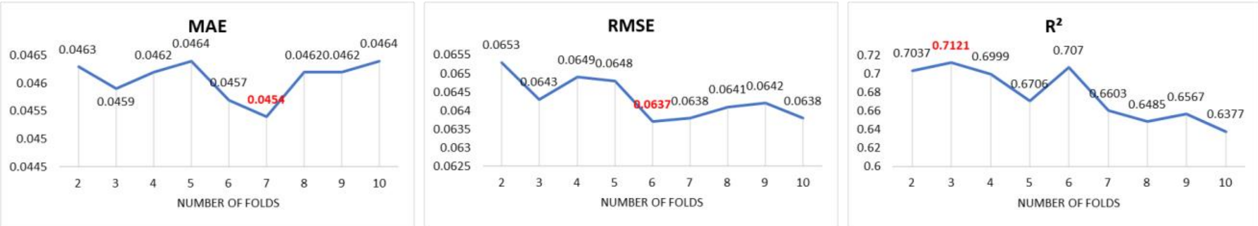


Figure 182. K-fold cross-validation for RT model with SVR and combination two

Combination three

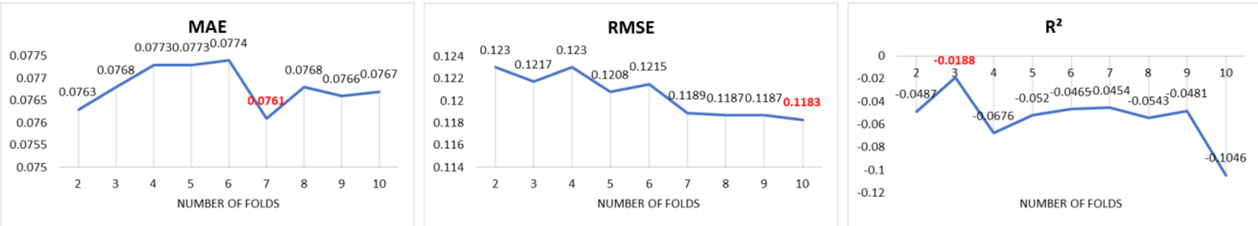


Figure 183. K-fold cross-validation for RT model with SVR and combination three

Artificial Neural Network

Combination one

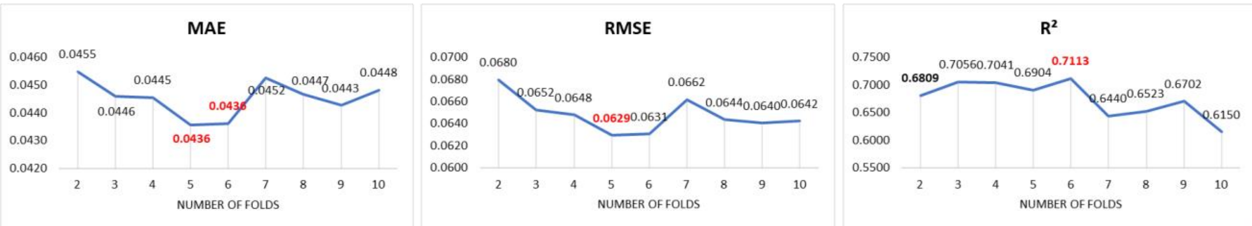


Figure 184. K-fold cross-validation for RT model with ANN and combination one

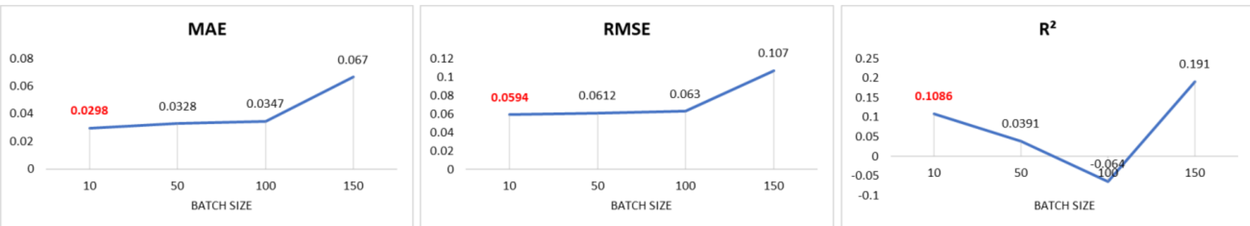


Figure 185. Batch Size for RT model with ANN and combination one

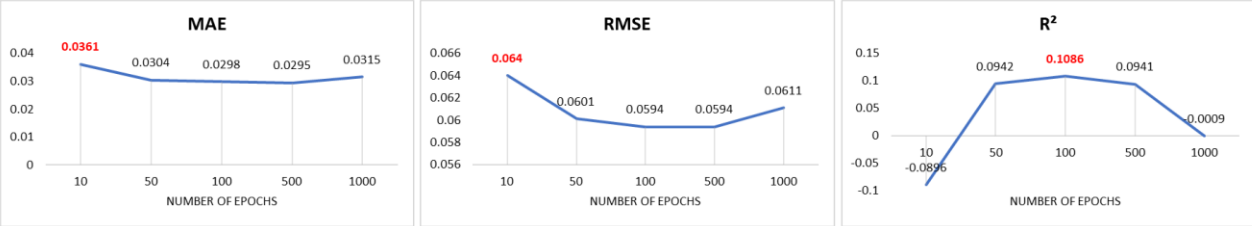


Figure 186. Epochs for RT model with ANN and combination one

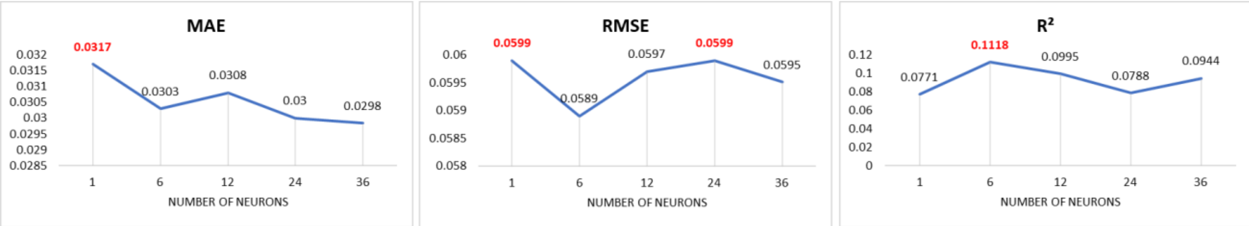


Figure 187. Neurons for RT model with ANN and combination one

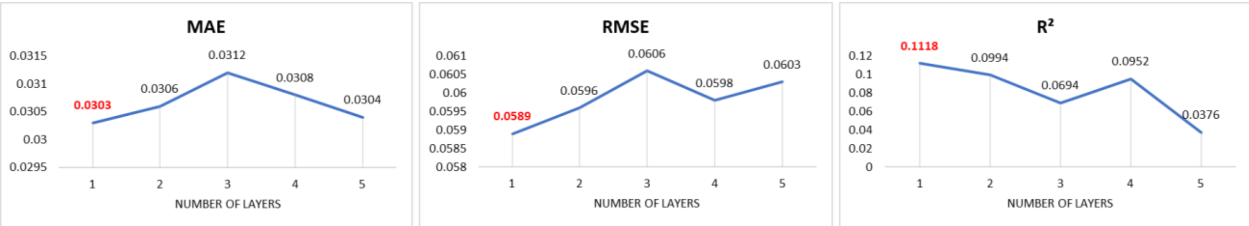


Figure 188. Hidden layers for RT model with ANN and combination one

Combination two

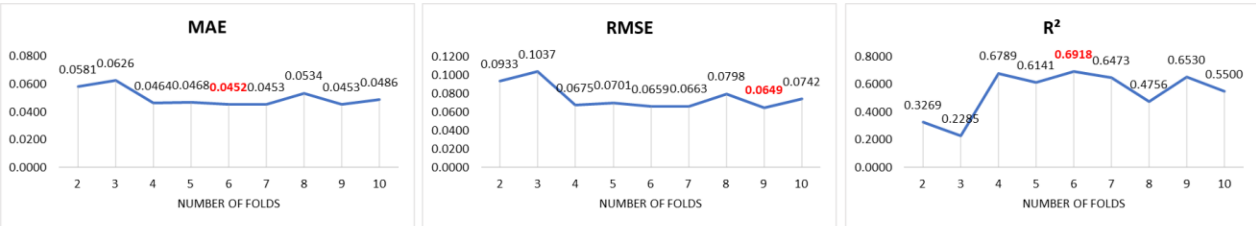


Figure 189. K-fold cross-validation for RT model with ANN and combination two

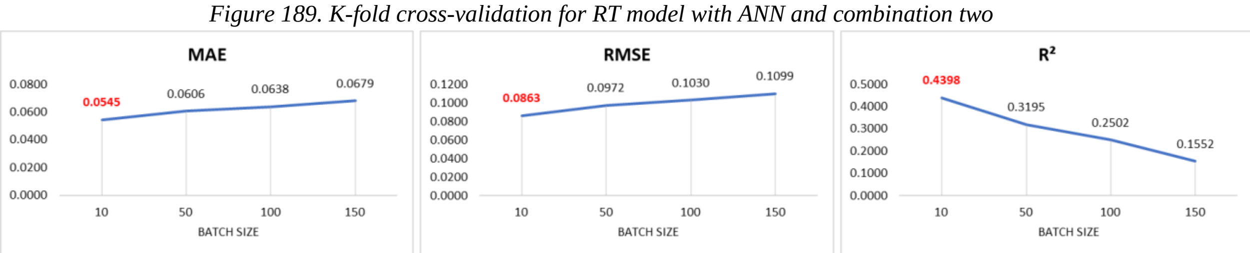


Figure 190. Batch Size for RT model with ANN and combination two

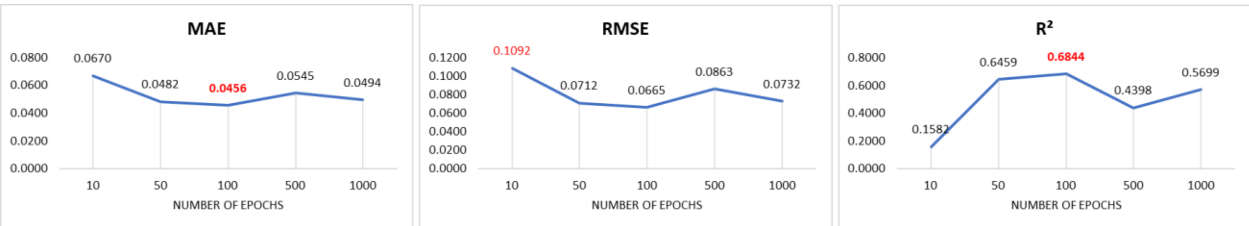


Figure 191. Epochs for RT model with ANN and combination two

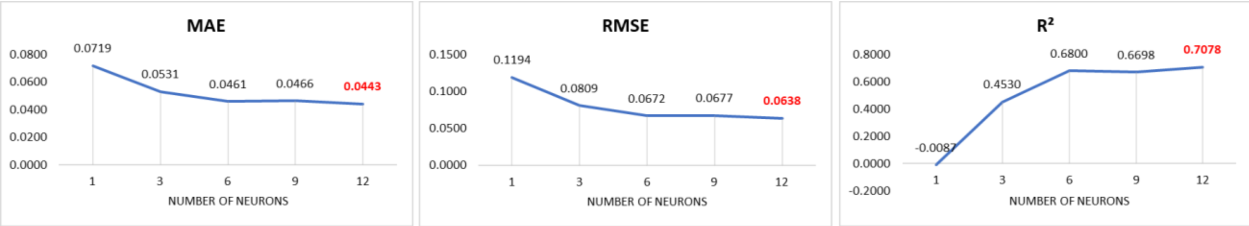


Figure 192. Neurons for RT model with ANN and combination two

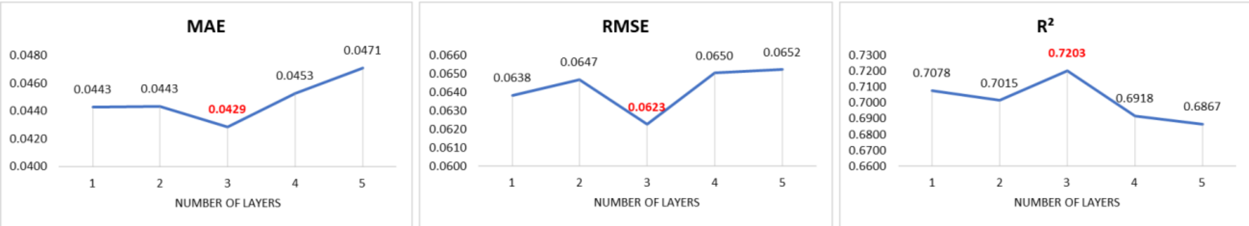


Figure 193. Hidden layers for RT model with ANN and combination two

Combination three

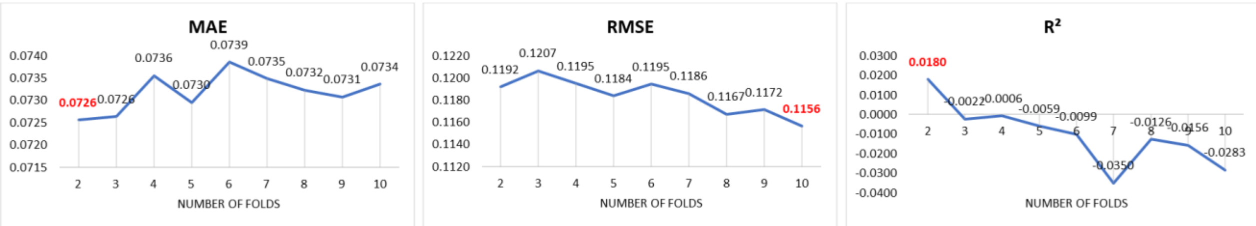


Figure 194. K-fold cross-validation for RT model with ANN and combination three

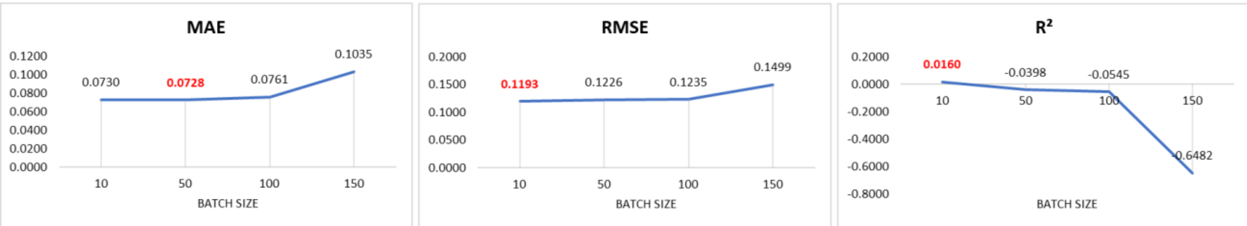


Figure 195. Batch Size for RT model with ANN and combination three

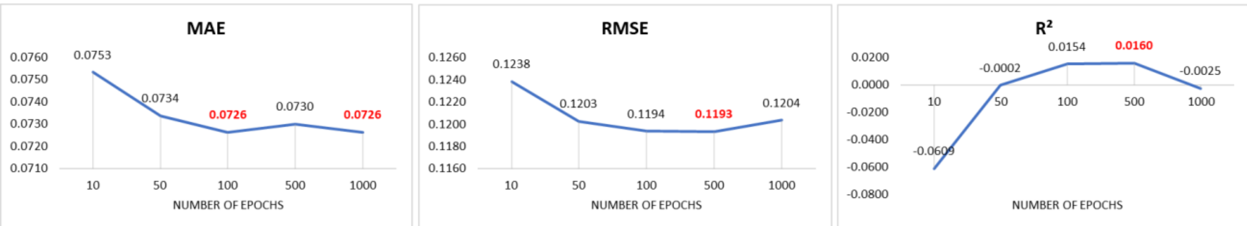


Figure 196. Epochs for RT model with ANN and combination three

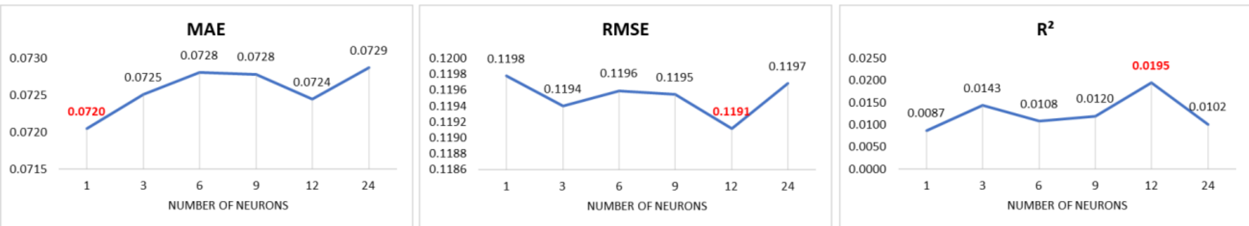


Figure 197. Neurons for RT model with ANN and combination three

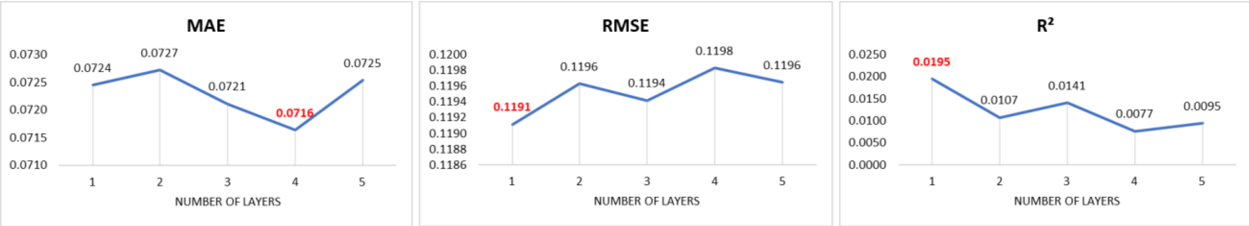


Figure 198. Hidden layers for RT model with ANN and combination three

Truck Travel Time

Multi Linear Regression

Combination one

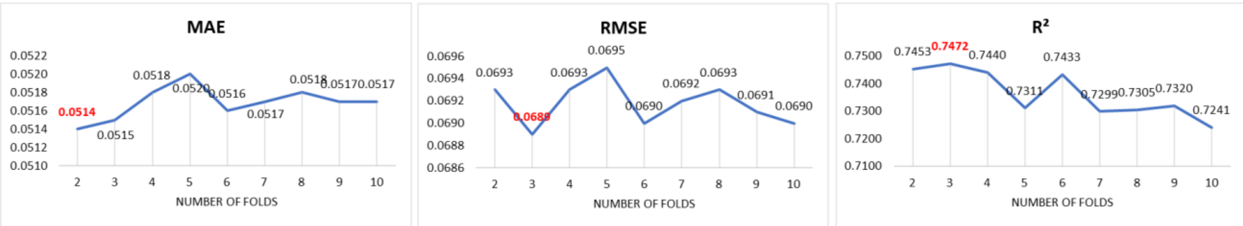


Figure 199. K-fold cross-validation for TTT model with MLR and combination one

Combination two

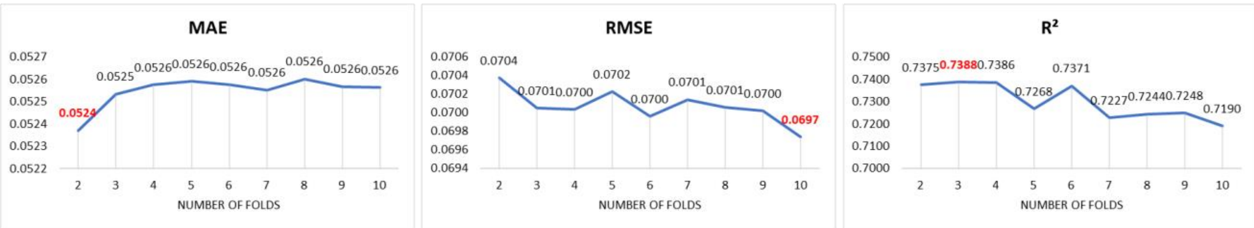


Figure 200. K-fold cross-validation for TTT model with MLR and combination two

Combination three

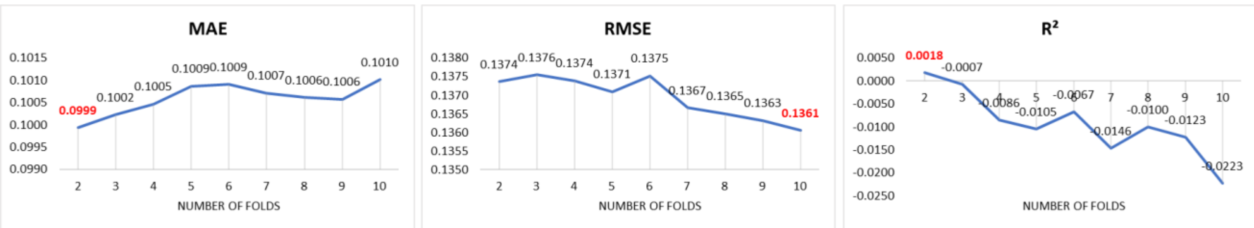


Figure 201. K-fold cross-validation for TTT model with MLR and combination three

Support Vector Regression

Combination one

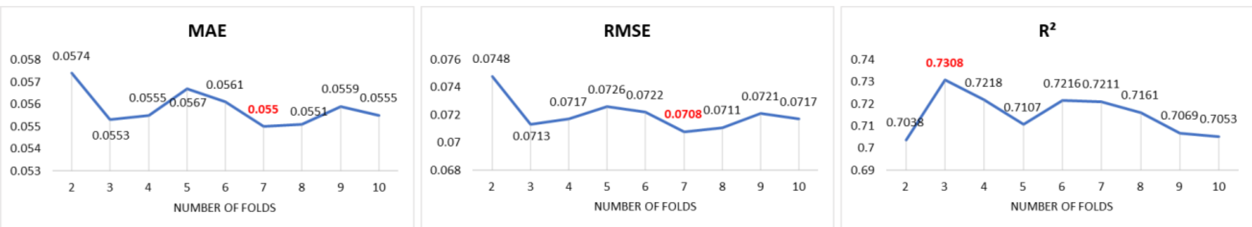


Figure 202. K-fold cross-validation for TTT model with SVR and combination one

Combination two

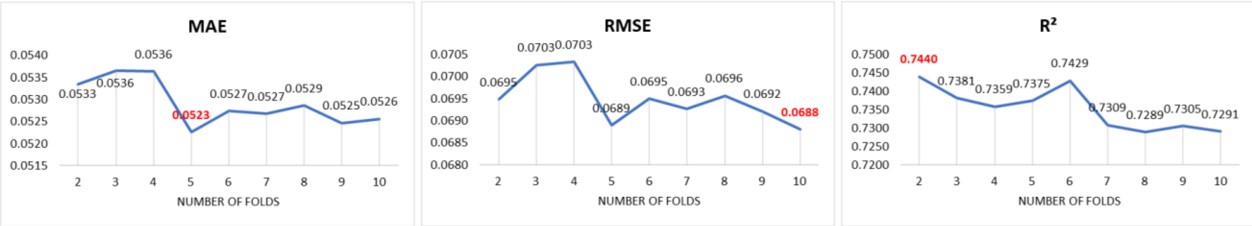


Figure 203. K-fold cross-validation for TTT model with SVR and combination two

Combination three

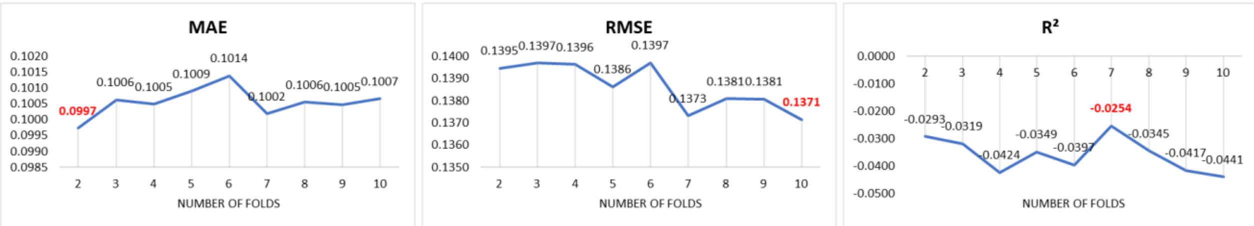


Figure 204. K-fold cross-validation for TTT model with SVR and combination three

Artificial Neural Network

Combination one

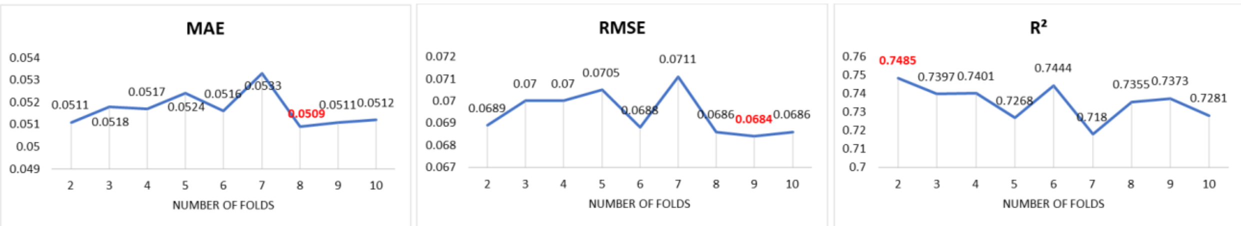


Figure 205. K-fold cross-validation for TTT model with ANN and combination one

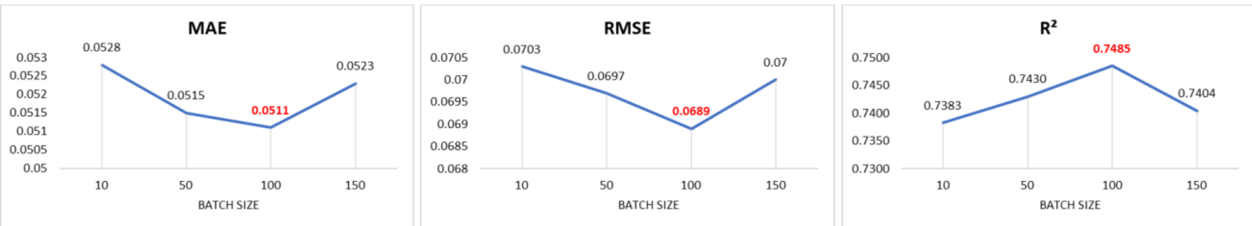


Figure 206. Batch Size for TTT model with ANN and combination one

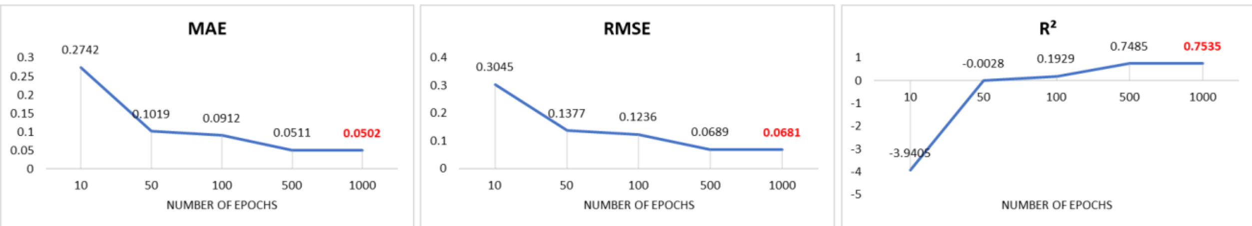


Figure 207. Epochs for TTT model with ANN and combination one

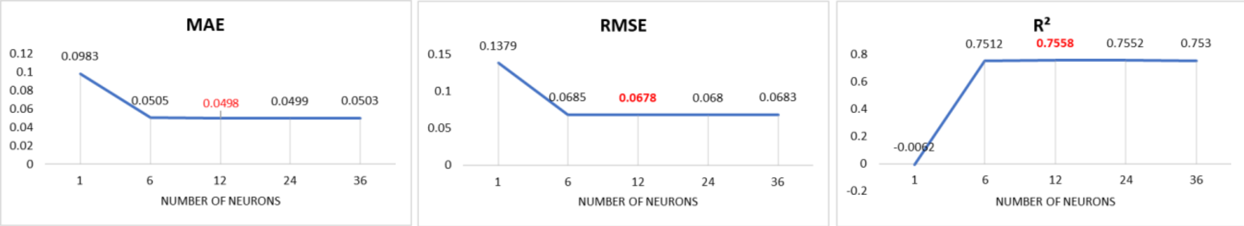


Figure 208. Neurons for TTT model with ANN and combination one

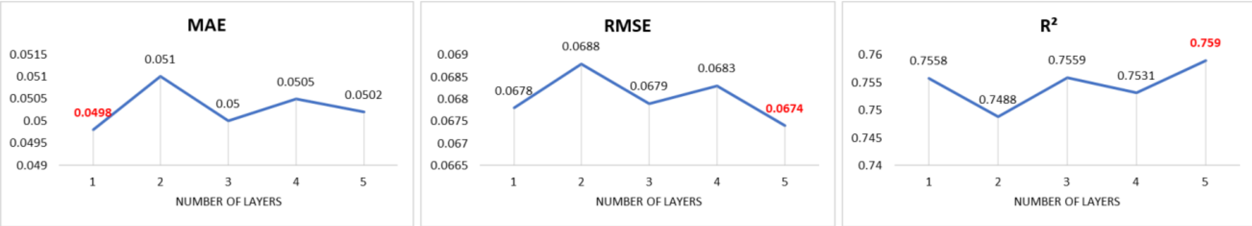


Figure 209. Hidden layers for TTT model with ANN and combination three

Combination two

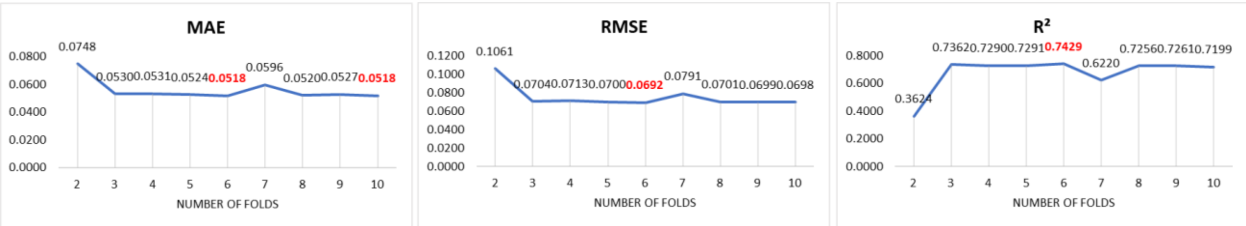


Figure 210. K-fold cross-validation for TTT model with ANN and combination two

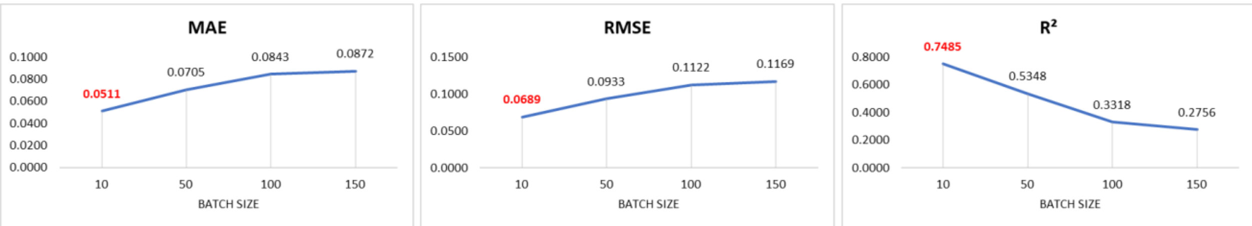


Figure 211. Batch Size for TTT model with ANN and combination two

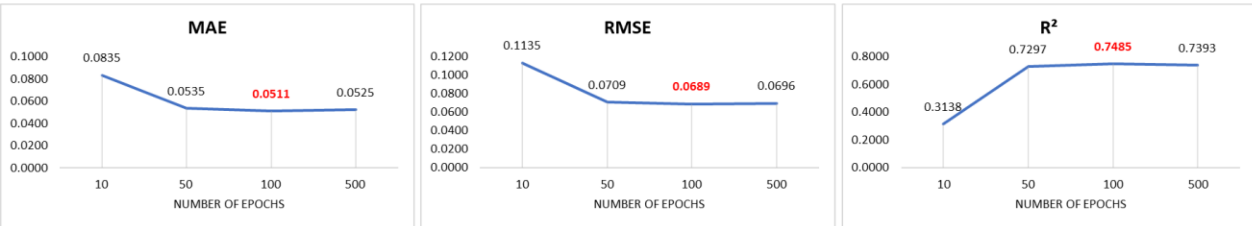


Figure 212. Epochs for TTT model with ANN and combination two

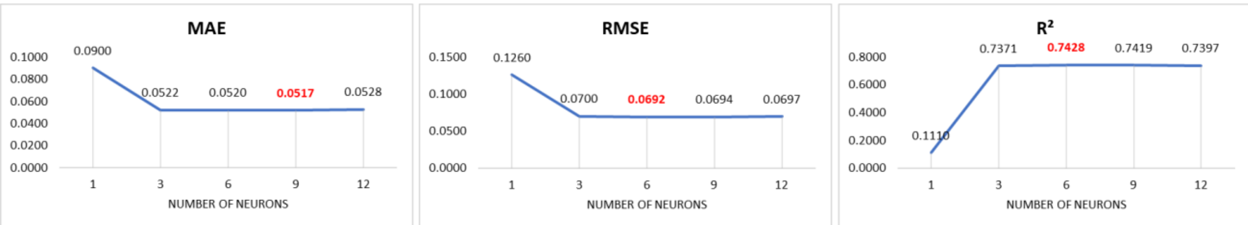


Figure 213. Neurons for TTT model with ANN and combination two

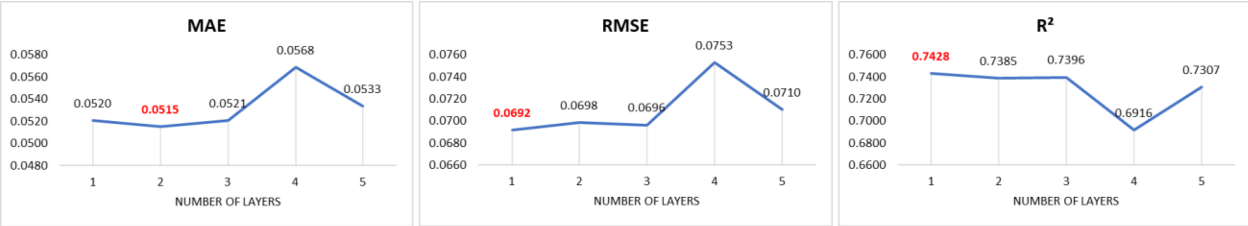


Figure 214. Hidden layers for TTT model with ANN and combination three

Combination three

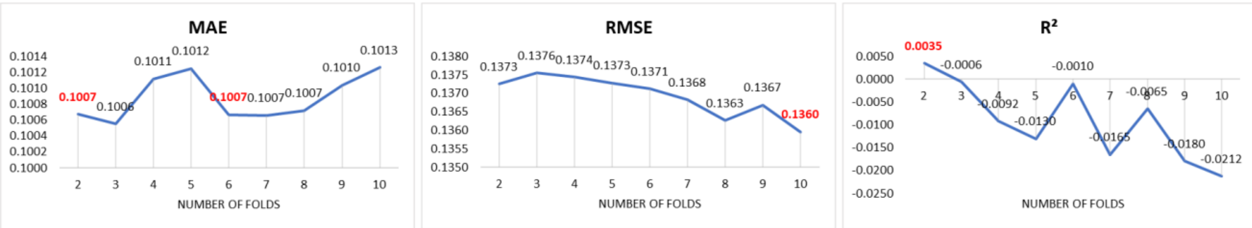


Figure 215. K-fold cross-validation for TTT model with ANN and combination three

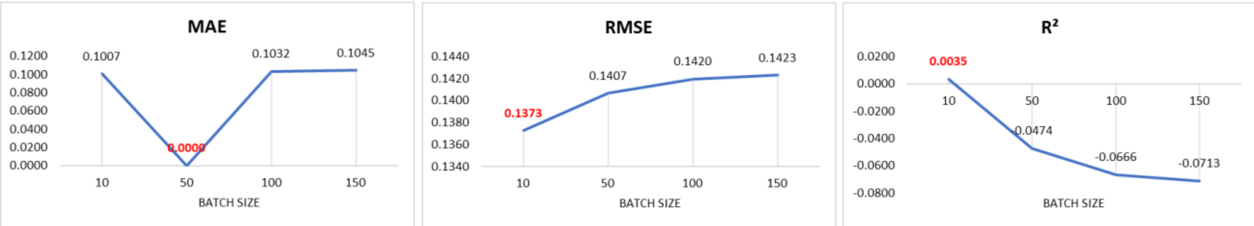


Figure 216. Batch Size for TTT model with ANN and combination three

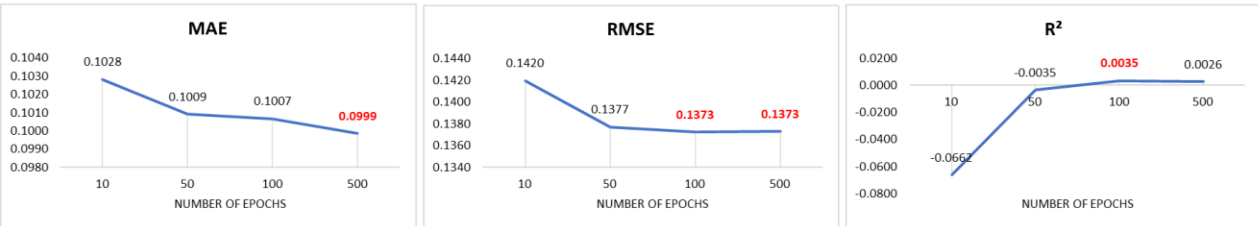


Figure 217. Epochs for TTT model with ANN and combination three

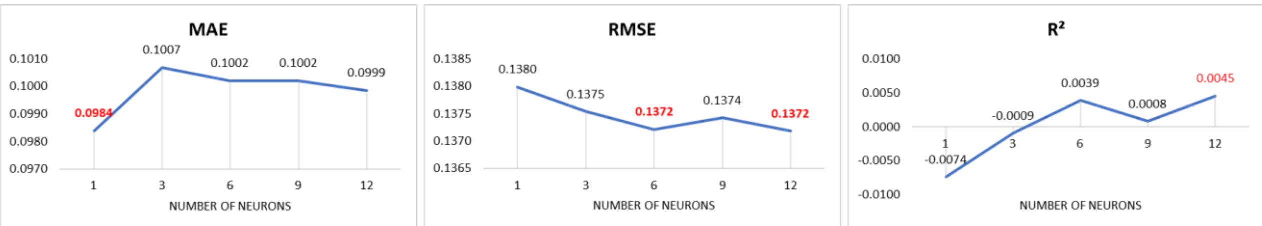


Figure 218. Neurons for TTT model with ANN and combination three

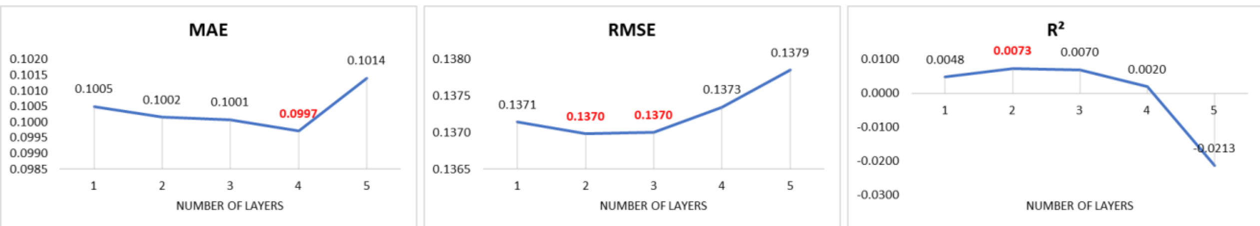


Figure 219. Hidden layers for TTT model with ANN and combination three

Truck Cycle Time

Multi Linear Regression

Combination one

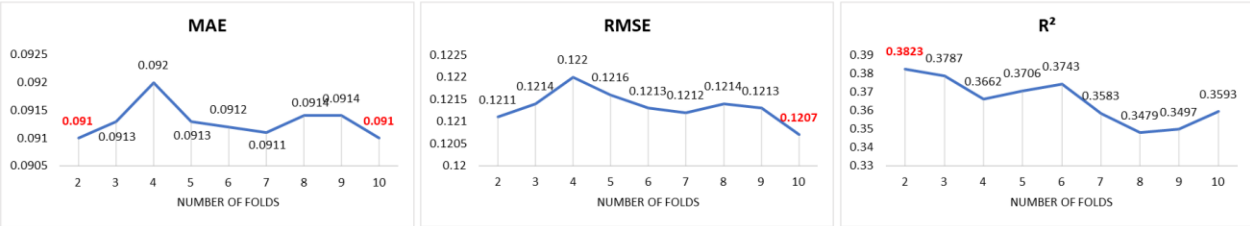


Figure 220. K-fold cross-validation for TCT model with MLR and combination one

Combination two

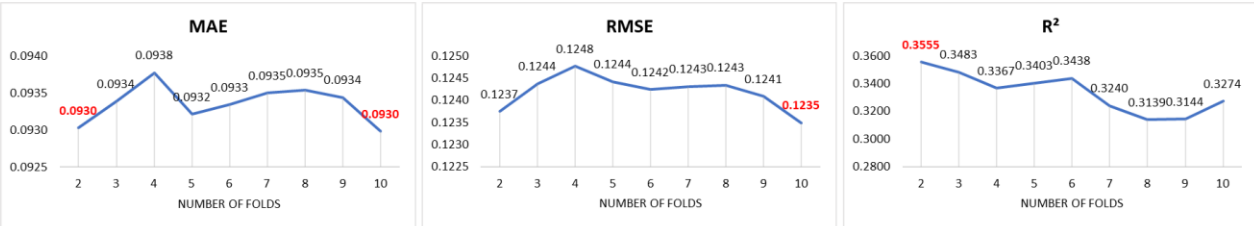


Figure 221. K-fold cross-validation for TCT model with MLR and combination two

Combination three

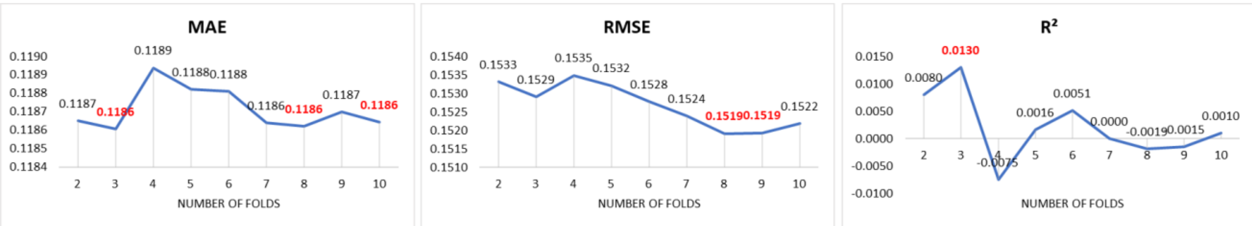


Figure 222. K-fold cross-validation for TCT model with MLR and combination three

Support Vector Regression

Combination one

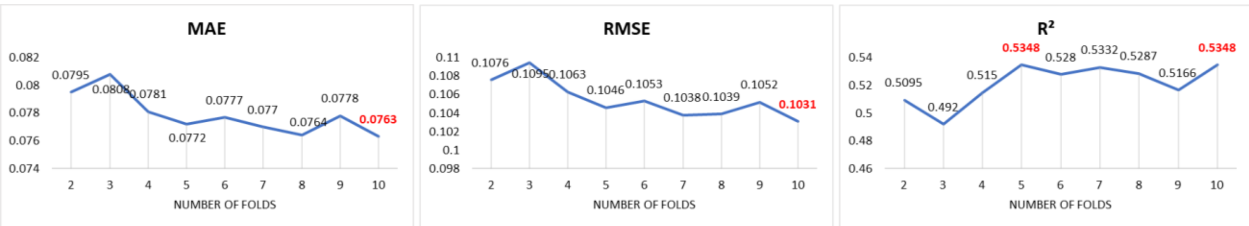


Figure 223. K-fold cross-validation for TCT model with SVR and combination one

Combination two

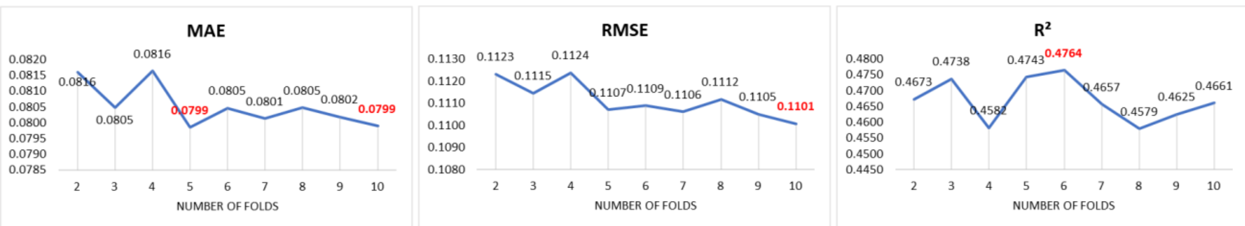


Figure 224. K-fold cross-validation for TCT model with SVR and combination two

Combination three

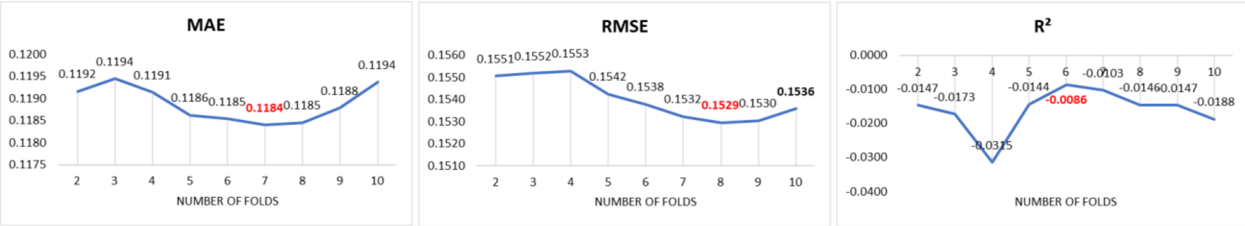


Figure 225. K-fold cross-validation for TCT model with SVR and combination three

Artificial Neural Network

Combination one

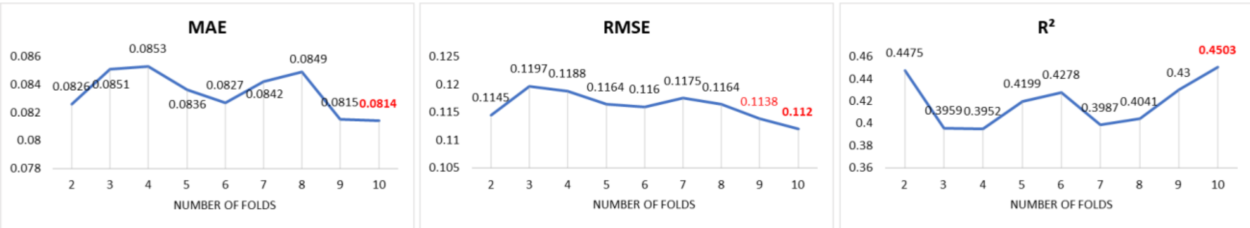


Figure 226. K-fold cross-validation for TCT model with ANN and combination one

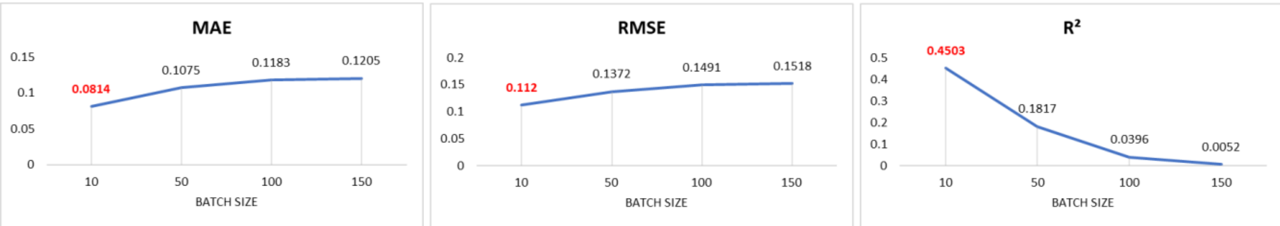


Figure 227. Batch Size for TCT model with ANN and combination one

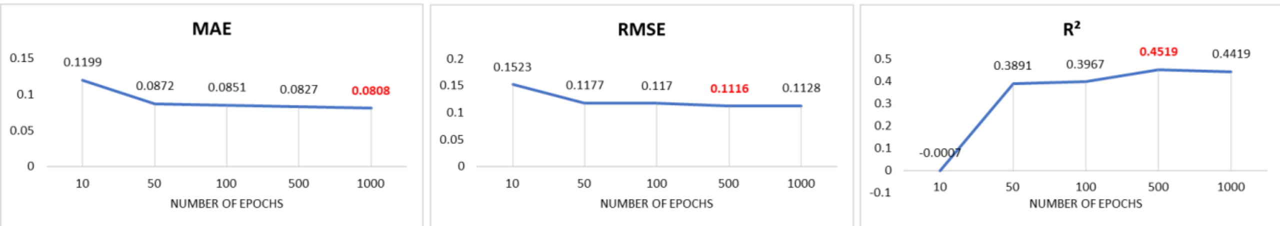


Figure 228. Epochs for TCT model with ANN and combination one

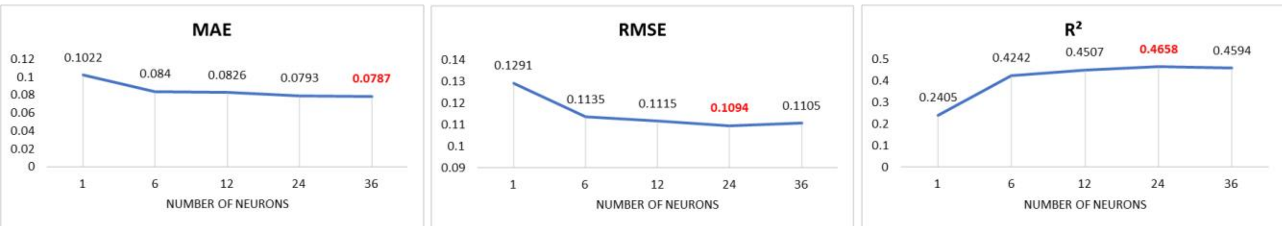


Figure 229. Neurons for TCT model with ANN and combination one

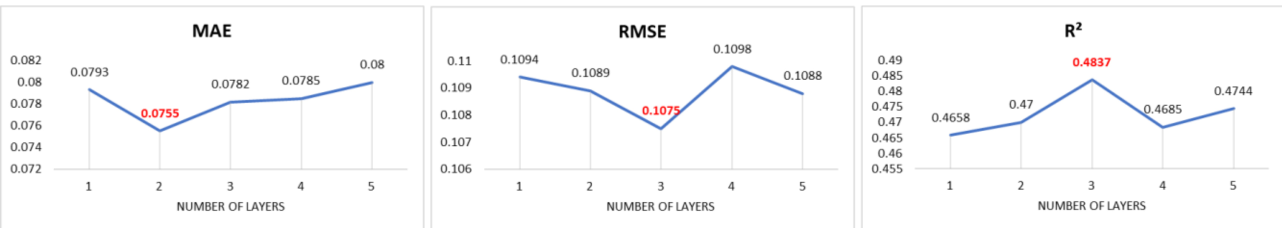


Figure 230. Hidden layers for TCT model with ANN and combination one

Combination two

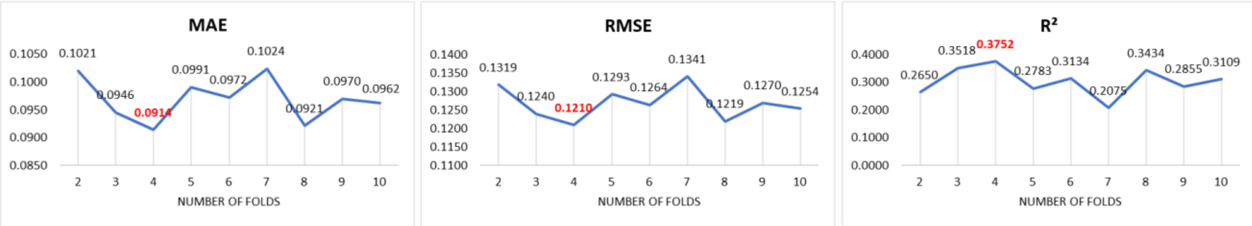


Figure 231. K-fold cross-validation for TCT model with ANN and combination two

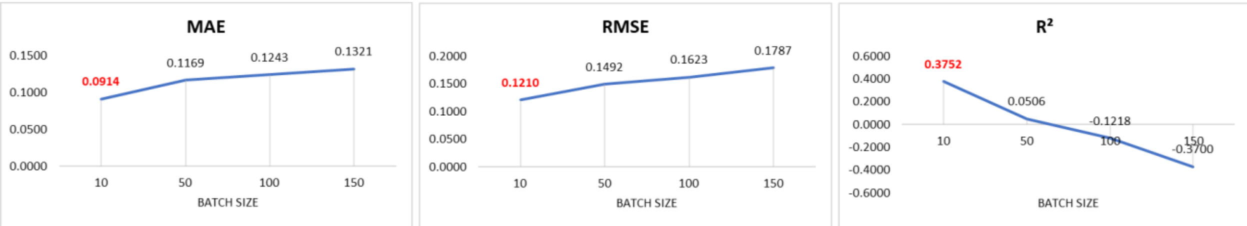


Figure 232. Batch Size for TCT model with ANN and combination two

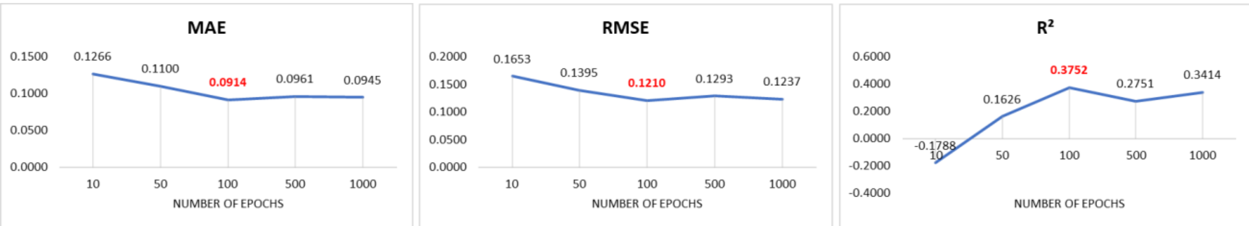


Figure 233. Epochs for TCT model with ANN and combination two

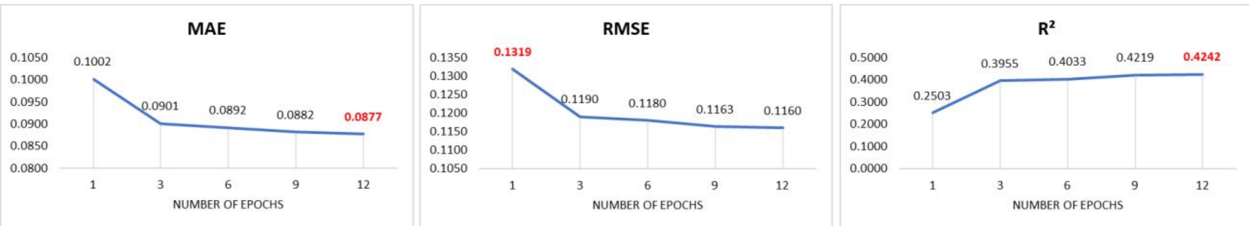


Figure 234. Neurons for TCT model with ANN and combination two

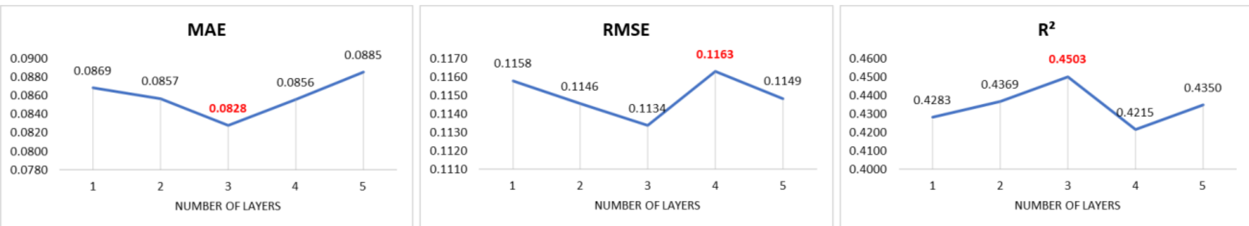


Figure 235. Hidden layers for TCT model with ANN and combination two

Combination three

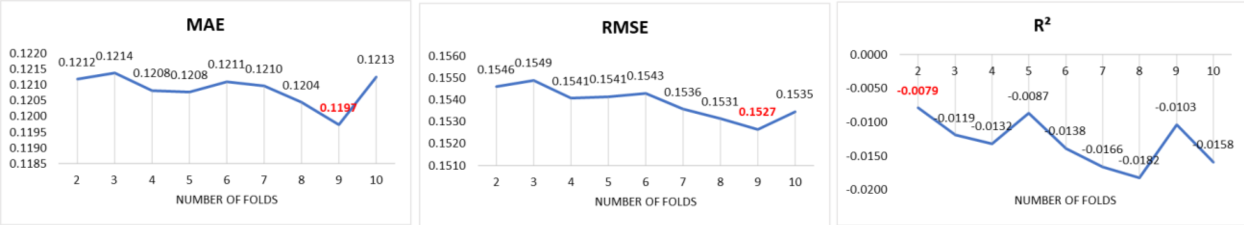


Figure 236. K-fold cross-validation for TCT model with ANN and combination three

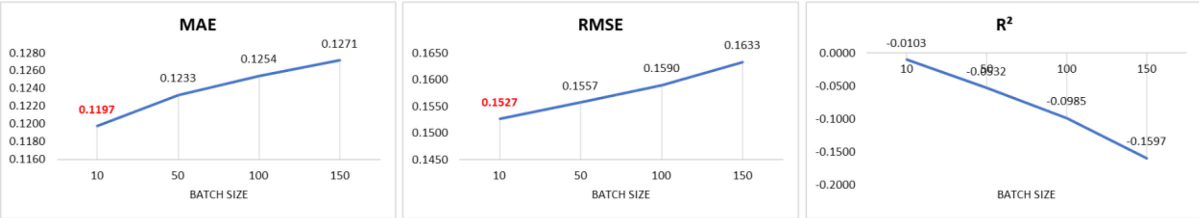


Figure 237. Batch Size for TCT model with ANN and combination two

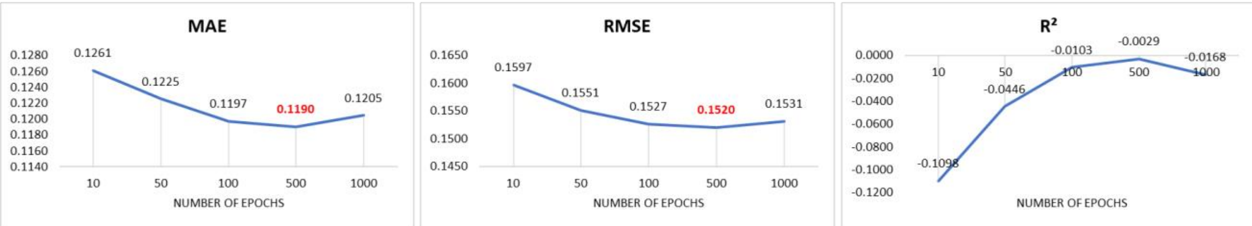


Figure 238. Epochs for TCT model with ANN and combination two

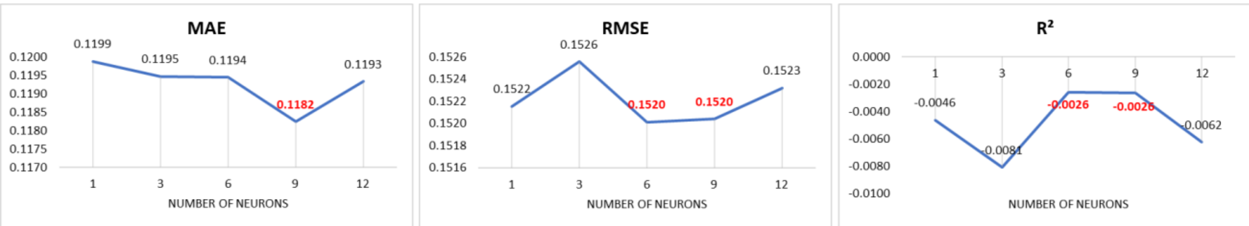


Figure 239. Neurons for TCT model with ANN and combination two

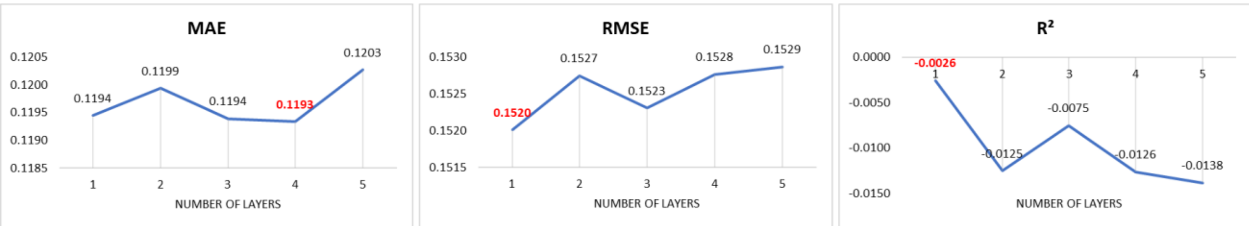


Figure 240. Hidden layers for TCT model with ANN and combination two

Appendix 3

This section aims to show the best prediction model of load time, haul time, unload time, return time, truck travel time, and truck cycle time models developed using MLR. Table 28 shows the coefficient value for each variable and intercept of each model that has been normalized. Thus, this model requires normalization of the input and denormalisation of the output to obtain the result in actual scale.

Table 28.MLR model result from automated data

Model	Temperature	Start Time Hour	Volume	Relative Humidity	Distance	Intercept
Load Time						
0.03	-0.11	0.02	-0.26	-0.075	-0.152	0.49
Haul Time						
-0.085	-	-	0.007	-	0.395	0.4
Unload Time						
-0.036	-	-	-0.052	-	-0.016	0.25
Return Time						
-0.01	0.031	-0.077	-0.165	-0.022	0.85	0.15
Truck Travel Time						
-0.045	-0.0001	-0.054	-0.123	-0.035	0.964	0.163
Truck Cycle Time						
-0.022	-0.098	-0.025	-0.336	-0.097	0.7522	0.452

Appendix 4

Interview

Yusuf Santoso – Project Manager PT. Abdi Sarana Nusa

Part I. Current Situation

1. a. How is currently TCT estimated?

We use simple math to estimate TCT and doing record manually

b. How is the result?

It is difficult because there are many combinations which is difficult to be done manually.

2. What are the effects that an early/late arrival of a truck?

It will affect the target, lost money because of the inefficiency. It also affects client trust because we promise to them to deliver on time.

3. What would the benefit of more accurate truck cycle time predictions?

Make easy by knowing the capacity and time needed. It also can help for better anticipation . and gain client trust.

4. What is the accuracy required for the truck cycle time to improve the efficiency in earthworks?

80% until 100%. However, gaining 100% is difficult because there are many factors involved. But the deviation of around 1 minute is acceptable.

Part II. Predictive Modelling

1. In your opinion, what is the most significant element in TCT?
    - ❑ Temperature
    - ❑ Relative Humidity
    - ❑ Start Time Hour
    - ❑ Distance
    - ❑ Volume
    - ❑ Model

Start Time Hour and distance is important for managing the drivers.

2. Can you think of other elements that might influence the result?

Road condition, type of material because the type of material affects the truck speed. If the material is not rigid, for instance, mud, the driver needs to be more careful in carrying the material.

Weather condition, rain or no rain.

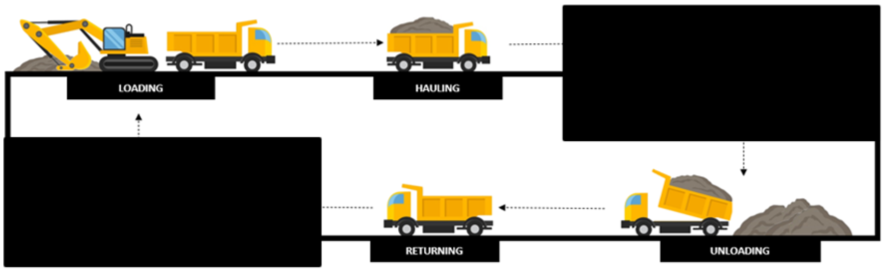
Driver behaviour and experience

Tires condition

3. The following table is the range of the predictive model.

	Distance (km)	Temperature (Celsius)	Relative Humidity (%)	Start Time Hour	Model	Volume (m^3)
mean	1.665419	15.907301	73.266452	12.353141		17.149157
min	0.600000	8.900000	49.330000	7.000000	ADT Type 1	1.140351
max	3.900000	18.200000	95.410000	17.000000	ADT Type 2	27.719298

Does this limitation cover most projects?
It will not cover most projects because the distance often more than 3.9 km. However, transporting overburdened material will be useful for a long term project.
4. Which activity in a truck cycle is important for you to know?



All of them are important. Load time and unload time are important because the location to unload sometimes is limited, and trucks need to wait.

5. Which of them that you want to be improved in terms of the accuracy level?

Load Time (LT)	Haul Time (HT)	Unload Time (UT)	Return Time (RT)	HT + RT	LT + HT +UT+RT
33%	31%	5.8%	78%	79%	56%

Load and Unloading time.

6. The result shows that it requires a short time to unload/load a big amount of material, volume and weight. In your opinion, What is the possible reason behind it?
It is possible because an excavator with a big material capacity tends to be faster than an excavator with a small capacity. The big excavator is also better at manoeuvring. Unloading material depends on the condition of the unloading area, whether it ready or not.

7.Can you think of other parties that might be interested in such an information system?
Perhaps, finance department. They need to estimate how many trucks that they need to buy for the company.

8. Do you have a recommendation for this research?
Add more vehicle type, for instance, Komatsu, XCMG, Hitachi, Dell.

Add road condition, whether flat or slope.

Anonymous- Logistic and Operation Analyst – Caterpillar Inc.

Part I. Current Situation

1. How is currently TCT estimated?

We have engineer review regularly to update the estimation, but it is confidential.

2. What are the effects that an early/late arrival of a truck?

It will impact the truck order. And, the company will ask the employee to overtime so the order can be achieved. So, we can't produce trucks in a big number because it will impact the warehouse.

3. What would the benefit of more accurate truck cycle time predictions?

It helps for better anticipation and gains client trust.

4. What is the accuracy required for the truck cycle time to improve the efficiency in earthworks?

I don't know.

Part II. Predictive Modelling

2. In your opinion, what is the most significant element in TCT?
   - ❑ Temperature
   - ❑ Relative Humidity
   - ❑ Start Time Hour
   - ❑ Distance
   - ❑ Volume
   - ❑ Model

Model. the project needs to use the right vehicle for achieving the

2. Can you think of other elements that might influence the result?

Type of excavator, site condition, drivers experience. If the soil needed to be excavated located in one location, it will be faster.

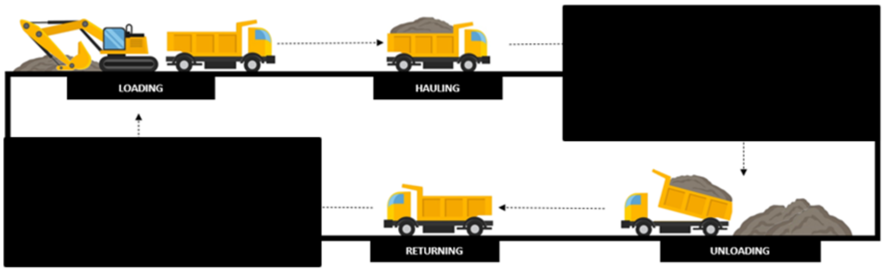
3. The following table is the range of the predictive model.

	Distance (km)	Temperature (Celsius)	Relative Humidity (%)	Start Time Hour	Model	Volume (m^3)
mean	1.665419	15.907301	73.266452	12.353141		17.149157
min	0.600000	8.900000	49.330000	7.000000	ADT Type 1	1.140351
max	3.900000	18.200000	95.410000	17.000000	ADT Type 2	27.719298

Does this limitation cover most projects?

I think yes, it covers most projects. But it depends on the type of project. this range might be useful for mining project

4. Which activity in a truck cycle is important for you to know?



All of them are important. For example, load time and unload time are important because the location to unload sometimes is limited, and trucks need to wait.

5. Which of them that you want to be improved in terms of the accuracy level?

Load Time (LT)	Haul Time (HT)	Unload Time (UT)	Return Time (RT)	HT + RT	LT + HT +UT+RT
33%	31%	5.8%	78%	79%	56%

Load Time because it relates to many types of equipment.

6. The result shows that it requires a short time to unload/load a big amount of material, volume and weight. In your opinion, What is the possible reason behind it?
Perhaps, the material is scattered, so it needs more time to load the material. Require time to process the dump material.

7.Can you think of other parties that might be interested in such an information system?
Dealer Company

8. Do you have a recommendation for this research?
Add excavator type.