International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

Using Machine Learning Methods to Predict Order Lead Times

Farhana Sethi*

Global Data & Analytics Business Intelligence - Quality & Governance Manager with Schlumberger Oilfield, Texas, Houston

Email: fsethi@slb.com, Tel.: +1-832-853-4059

Abstract

The precise prediction of end-to-end Parts shipments delivery lead time (LT) considerably influences the efficiency and quality of manufacture planning and job scheduling in Oil and gas industry. Lead time transparency and predicting precise lead times allow Oil and gas industry to reduce operating expenses, enhance capital, upturn revenues, and improve their viable advantages. Clients will be able to better assign resources well and reduce the risk through conviction in their product and services allocation. The lead time prediction using machine learning algorithm can overall improve the job scheduling, improve service levels, gain efficiency, lead to reduce cost, and improves customer satisfaction. This paper describes the algorithm and techniques to execute the research using Machine Learning Methods to Predict Order Lead times. We are going to share the prediction accuracy results using different algorithm and share the sensitivity analysis results. The algorithm used to train the model consumed historical data from organization data using logistic application and various variables which are specific to the Supply chain in the Oil and gas industry.

Keywords: Machine Learning; Lead time; Random Forest; Logistic; Production.

1. Introduction

Predictive data analytics contains a variety of statistical methods from data mining, predictive modelling, and ML that analyze historical facts to make predictions about prospects or unidentified occasions.

---

* Corresponding author.

"CRISP-DM breaks the process of data mining into six major phases:1) Business Understanding; 2) Data Understanding; 3) Data Preparation; 4) Modeling; 5) Evaluation; 6) Deployment" [1]. The fourth stage (modeling) is where machine learning (ML) algorithms are engaged to build prediction for Lead time. In particular, ML is defined as automated computational method that "learns" and extract information and patterns directly from (historical) data. There are four approaches, 1) Supervised, 2) Unsupervised, 3) Semi-Supervised, and 4) Reinforcement Machine Learning.

- Supervised ML – Is a learning task with full set of labeled data while training an algorithm. In other words, it accepts that training instances are classified or labeled (learning affiliation between a set of descriptive features and a target feature). In Supervised ML, the model needs to be trained on a classified dataset that means we have both raw input data as well as its outcomes. We divided our data into a training dataset and test dataset, where the training dataset is used to train our system, whereas the test dataset turns as new data for predicting results or to see the accuracy of our model. Supervised ML is relatively a simpler, highly accurate, and trustworthy method. There are three key areas where supervised learning is beneficial: Classification Techniques, Regression, and Forecasting. Classification techniques uses the algorithm to predict a discrete value, focused on predicting a qualitative reaction by analyzing data and identifying patterns. On the other hand, regression technique used continuous data. The technique classically used in predicting, forecasting, and finding relationships amongst quantitative data. While Forecasting is the process of making predictions about the prospect based on the past and present data. It is most commonly used to analyze trends.

- Unsupervised ML- Concerns the analysis of unclassified examples. In unsupervised learning, a deep learning model is offered a dataset without categorical commands on what to do with it. The training dataset in this approach is a collection of instances without a specific desired outcome or an accurate answer. Unsupervised learning is computationally compound, less accurate, and reliable method.
- Semi-Supervised ML (SSL) - It is expected that there are also unlabeled data available at the time of training in addition to the labeled data. The objective of SSL methods is to excerpt information from the unlabeled data that could ease learning a discriminative model with greater performance.
- Reinforcement Machine Learning - Is a category of machine learning techniques that enables an agent (Artificial Intelligent agent) to learn in a collaborating environment by trial and error using response from its own activities and experiences.

"Regression technique in supervised machine learning predicts a single and continuous target output value using training data" [2]. In our scenario, modeling the association between the continuous variable (e.g. lead-time) and one or more predictors (For example, supplier origin, Buyer destination, Order type, etc.) using a linear function is the most suitable approach. One of the major strengths of using regression algorithm is that the Outputs always have a probabilistic analysis and can be normalized to avoid overfitting. However, the Weaknesses of Logistic regression may underperform when there are numerous or nonlinear decision limitations. This method is not malleable, so it does not capture more complex connections. In this paper, we emphasis on supervised machine learning to predict the lead-time. The paper is organized as follows: The first unit presents the advanced approaches in lead-time prediction using regression algorithms. Second unit provides a methodology

for handling regression jobs which is demonstrated in the context of Oil and gas industry. Finally, we discourse on the main findings and identify future research capabilities.

2. Background and concept

A lead time is the latency between the initiation and completion of a process. For example, the lead time between the placement of an order and delivery of new cars by a given manufacturer might be between 2 weeks and 6 months, depending on various particularities. One business dictionary defines "manufacturing lead time" as the total time required to manufacture an item, including order preparation time, queue time, setup time, run time, move time, inspection time, and put-away time. For make-to-order products, it is the time between release of an order and the production and shipment that fulfill that order. For make-to-stock products, it is the time taken from the release of an order to production and receipt into finished goods inventory. This research focus on make-to-stock products. Lead time in inventory management is the lapse in time between when an order is placed to replenish inventory and when the order is received. Lead time affects the amount of stock a company needs to hold at any point in time. When considering the total amount of time for a purchase order to be delivered from a supplier, factor in the time taken for the supplier to accept and process the order. Lead time directly affects your total inventory levels. The longer your lead time the more stock you will need to hold in your inventory. Longer lead times make deliveries more unpredictable and force a company to rely heavily on demand forecasts to make orders. Once you have calculated your lead time, the next step is to employ corrective measures to reduce it. While there have been improvements to shipping and freight services in recent years, there are other factors that affect lead times. Orders may take time to process and be approved within your business, your suppliers will then need to place orders of their own for the materials to create your products, and there might be delays due to checks done at the ports by customs. There are different trade control compliance and regulations for each country that needs to be factored in an international organization. Traditionally, lead time calculation could be done using all those factors into consideration. For all of these factors add to your lead time. Hence, there are several factors required to be considered to calculate lead time in a specific industry. However, with the advancement of the technology, predictive analysis techniques can be used to predict the lead time based on the products and supplier/manufacturer.

2.1. Related work

Many research have focused on finding the right design, implementation, and evaluation of the lead time for various industries. In the research paper, Lukas and his colleagues outlines the most relevant research that is very close to this paper. They first discuss knowledge discovery approaches in production planning and control (PPC) then explore state-of-the-art regression-based reproaches in lead time prediction. “Statistical learning, data mining or knowledge discovery was first defined in 1989 as a new intelligent tool for extracting useful information and knowledge from different databases. The main purpose of production planning and control (PPC) is to establish routes and schedules for the work that will ensure the optimum utilization of materials, workers, and machines” [3]. Cheng and his colleagues revised the relevant papers since 2010 and discussed the typical knowledge mining techniques in production management. According to this survey, “The four most reflected typical application areas are advanced planning and scheduling, quality improvement, fault diagnosis

and defect analysis. A fifth category was defined, in which flow time/cycle time prediction could be found – among life and yield prediction. PPC was identified as a research gap already in 2009, and the review in 2017 has revealed just a few applications in this passed 9 years" [4]. Consequently, more attention from the research community would be needed to data mining in PPC. Ensuring the operation of the plant in accordance with these plans.

3. Methodology

From CRISP-DM model, the first three phases related to data processing as described in Figure 1.The unit ML algorithm is the modeling phase, and the evaluation is done in the Results section, where pertinent landscapes have been picked and the accurateness of prediction model has been calculated.

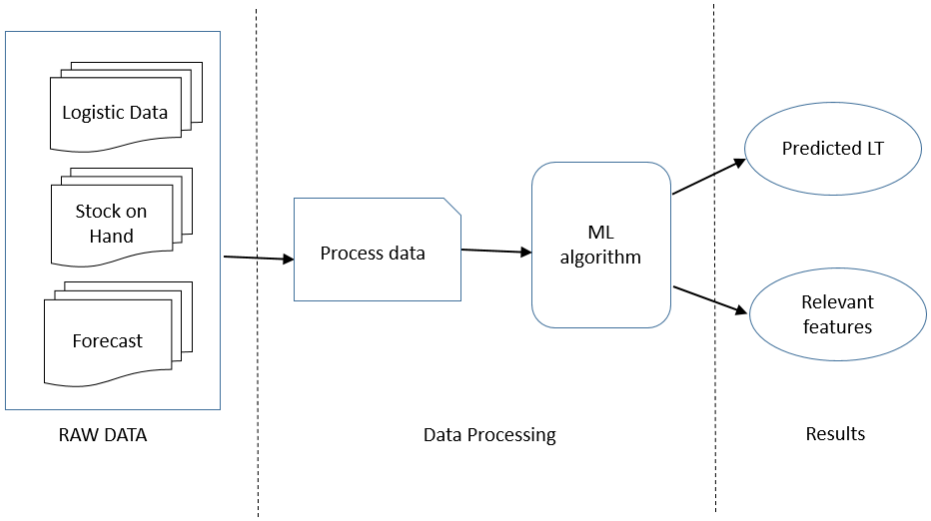

Figure 1: Data Analysis and Processing Steps

As specified above, the case study has been led in the Oil and gas industry. A distinct sort of this industry is that most of the products Lead times are fabricated in manifold layers. Many production plants have cross-used products among various product lines with the option of ordering Parts from Distribution centers, also the plants are organized in operation of multiple field jobs. This leads to a job driven production system, where products have to execute on the multiple jobs numerous times in order to achieve optimum utilization of logistics. Moreover, Industry also has a trend to forecast the future jobs and hence the usage of the product to complete the job in the same location. Having said that, many times the stock can reside and maintained in the location warehouse after completing the required job while waiting for the next assigned job. Due to the intricacy defined above, the Oil and gas industry conventionally invested a lot in advanced IT systems to build Stock replenishment tools for the parts stock on hand visibility, resulting in a large amount of, detailed data dimensions available for Stocks availability. The data that was used for the creation of the model was a dataset used by the logistic data warehouse that enclosed half a million rows and 50 plus columns, involving a combination of numerous data sources that the company has at its usage. This data encompassed various information such as the type of product, different manufacturers, various suppliers from different locations, the

buyer of the product, forecast of the consumption, and various dates and coordinates related to the order and delivery. Every line has a unique data set ID and a dimension of order lead time in days, which became the dependent variables. The 1st phase in the supervised learning data process stage is to collect labeled training data. The label is the output and delivers feedback for the algorithm. Provided sufficient data is obtainable in the raw data collection phase, the next stage is to divide this labeled data into three sets: training, testing, and validation to generate the ML algorithm. The algorithm uses the training set to regulate the model to reduce the error.

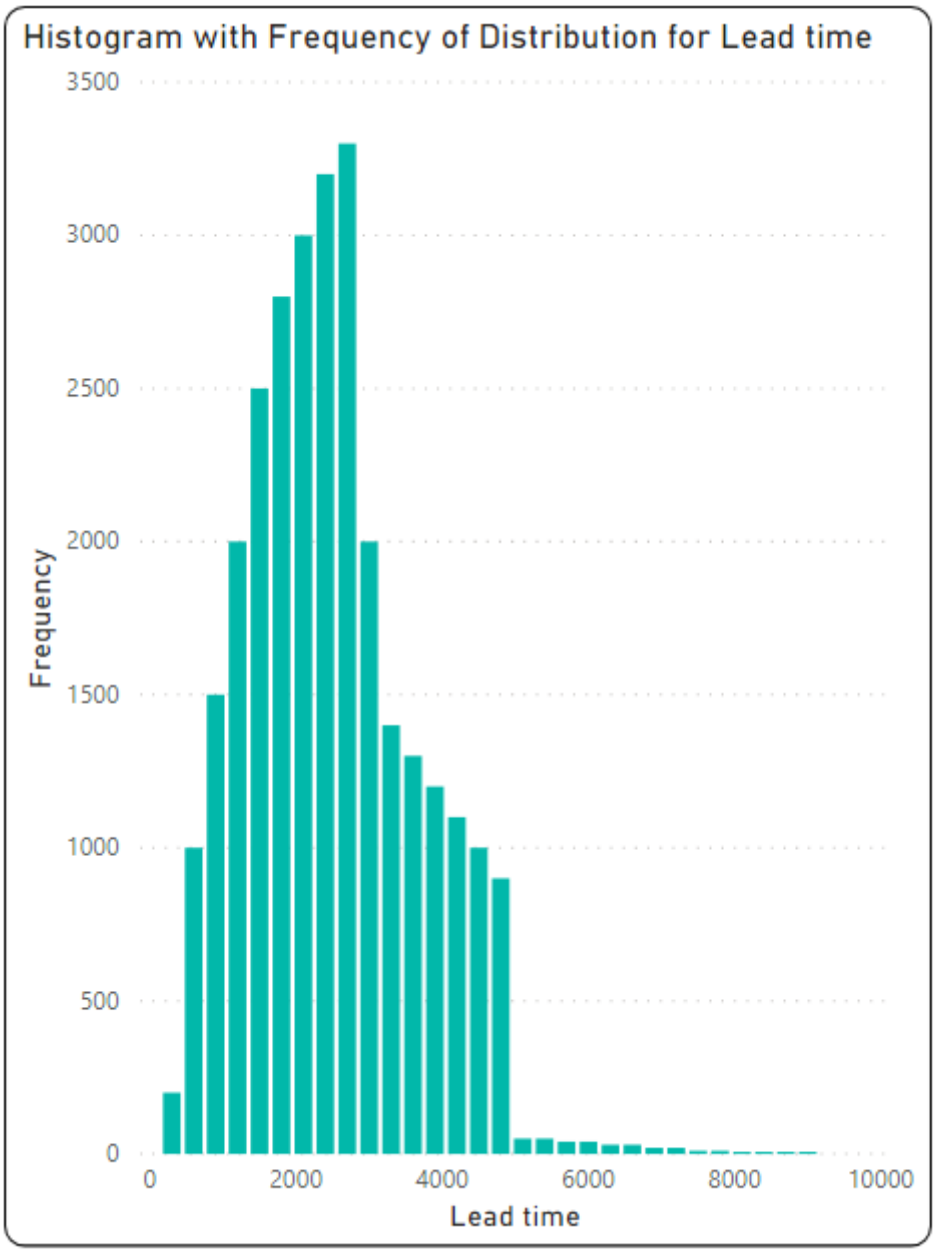

Figure 2: Histogram of the lead time prediction analysis

The validation set is severe from the training set and allows one to unconventionally measure the advancement of the learning algorithm. This measure can be used to define a cutoff point in the training algorithm to

steadiness the accuracy of the learned model against overfitting. The test set is the final set and it is intended to be used only when the model has been set up to be the finest on the validation set. This set offers a real domain assessment of the model performance on unknown data. Test data is a kind of final check for a model which has learned its training data effectively and can generalize to new data during data processing. A Histogram below in Figure 2 provides a visual representation of the data distribution of mathematical data by showing the number of data points that fall within a definite range of values. Histograms can exhibition a large amount of data and the frequency of the data points. The median and distribution of the data can be comprehended visually. In addition, it can show any outliers or gaps in the data. **Lead time** on X-axis: **The frequency** on Y-axis and the bar is the altitude indicates the number of times that the values unused.

3.1. Machine learning algorithms

In this section, the different machine learning algorithms are described. Then, we show different error measures. Next, the accuracy of the verified ML procedures based on various error measures is presented. Finally, a sensitivity analysis of outcomes of the significant variables of the models are used to predict the model. Machine learning algorithms are mathematical programs that can learn from data and develop from knowledge, without human intervention. To build the supervised learning model, each training should include a set of input features and a definite output value. . In case of regression, the importance of a continuous variable is to be predicted. The utmost major regression methods are the simply interpretable linear regression models (LM), which assume a just about linear relationship among the variables. Below algorithm method have been used to train the model.

- **Linear Regression** is a statistical approach that models the relationship between input features and output. The input features are called the independent variables, and the output is called a dependent variable
- **Ridge Regression** is the most ordinarily used regression algorithm to estimate an answer for an equation with no distinctive solution. This type of problem is widespread in machine learning tasks, where the "finest" solution must be elected using limited data.
- **Lasso Regression** (**Least Absolute Shrinkage and Selection Operator**; also **Lasso** or **LASSO**) is a regression analysis method that executes two variable collection and regularization in order to progress the prediction accuracy.
- **Random Forest (RF)** RF makes predictions by combining the results from many singular decision trees - so we call them a forest of decision trees, as described in Figure 3. Since RF combines several models, it falls under the group of ensemble learning.

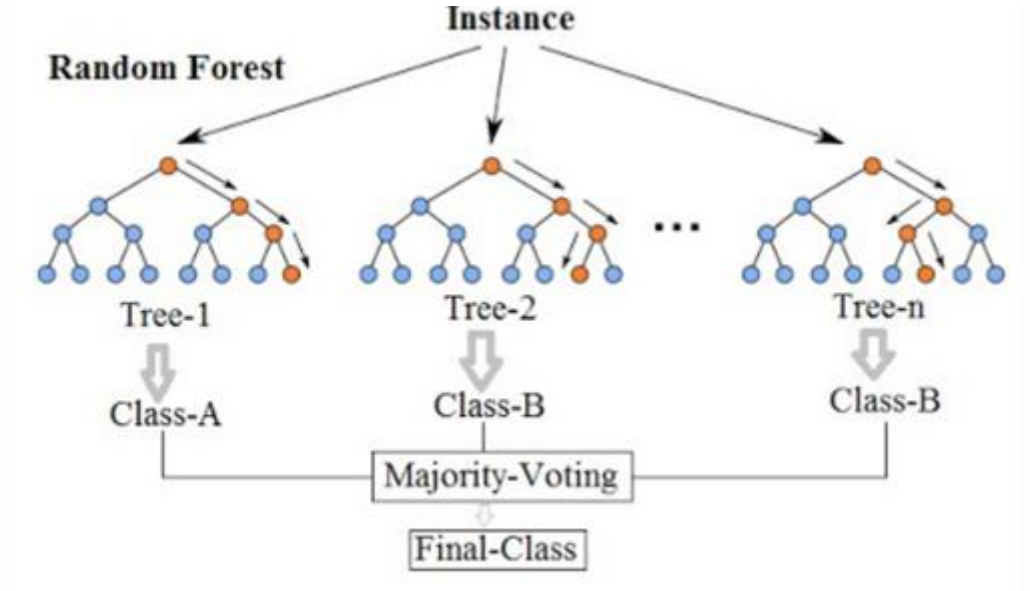

Figure 3: Random Forest technique

- **Support Vector Machines** (**SVM)** is a supervised machine learning method that is widely used in design, recognition, and classification of hitches, when data has surely two classes.
- **Multivariate adaptive regression spline (MARS)** is a nonparametric regression technique and can be perceived as an addition of linear models.
- **K-Nearest Neighbors (KNN)** procedure uses the whole data set as the training set, instead of piercing the data set into a training set and test set.
- **Artificial Neural networks (ANN)** or neural networks are computational procedures. It envisioned to pretend the activities of organic structures composed of "neurons". ANNs are computational models enthused by an animal's central nervous system. It is capable of machine learning as well as pattern recognition. These presented as systems of interconnected "neurons" which can compute values from inputs. Artificial Neural network is typically organized in 3 layers; Input layer, Hidden layer, and Output layer. ANNs are considered as simple mathematical models to enhance existing data analysis technologies. Although it is not comparable with the power of the human brain, it is the basic building block of the Artificial intelligence.

3.2. Lead time prediction with regression: state-of-the-art methods

In the present paper, lead time –as one of the most important control parameter and target figures of PPC– is analyzed and predicted with the help of different ML algorithms. As described in table: 1 below, we have applied all the above-mentioned model techniques and achieved the resultant accuracy.

Table 1: Precision of the established ML algorithms

	LM	RIDGE	Lasso	RF	SVM	MARS	KNN	ANN
MASE	30.51	40.5	40.23	30.42	27.81	38.835	39.6	48.06
MAE	357.3	459	457.2	351	380.7	439.2	453.6	481.5
MSE	332472.6	516168	515645.1	324702	450623.7	462274.2	49940.1	592966.8
RMSE	546.3	681.3	680.4	540	636.3	644.4	670.5	693.9
NRMSE	19.48	22.27	22.27	19.3	21.343	21.46	22	23.62

Performance metrics (error measures) are dynamic mechanisms of the assessment frameworks in several fields. “Performance metrics (error measures) are vital components of the evaluation frameworks in various fields” [5]. Deviation is useful for accuracy of continuous variables. Deviation is the difference between two or more values. E.g. prediction and actual, thus a measure of accuracy or performance. Deviation is also frequently simply called "error". There are 5 major error measures as follows:

- **Mean Absolute Scaled Error (MASE):** MASE is good if your accuracy must be comparable with other accuracies based on different datasets.
- **Mean absolute error (MAE):** MAE is one of the simplest measure of deviation.
- **Mean square error (MSE):** Square-based deviations are sensitive to outliers. Square-based deviations are used because larger errors have big effect on the score while small errors have little effect. They should be more useful when large errors are particularly undesirable.
- **Root MSE (RMSE):** Root-mean-square error (RMSE) is by far the most common deviation.
- **Normalized Root-Mean-Square Error (NRMSE):** NRMSE is useful if you are comparing accuracies of two different datasets. The original RMSE is scale-dependent and is similar to MASE. The concluding model recommended for lead time prediction in our case is the Random Forest model as described in below Table: 2. Model running take about 35 seconds, With NRMSE of 19.3 accuracy is reached.

Table 2: Explanation of sensitivity analysis outcomes of the eight best significant variables of the model

Feature	Feature Description	RF	LM
Destination Country	Buyer destination country and location	2.4	2.6
SOH	If stock is available to ship readily	6.5	4.9
Part Number	Ordered part	4.8	3.5
Order Type	If Item is enrouting via Distribution services center or Purchase order is directly for the supplier	3.4	2.8
Shipment Status	Various status (e.g. AT CUSTOMS, CUSTOMS CLEARED,ENROUTE TO DELIVERY LOCATION)	4.7	5.5
RDD	Requested Delivery Date	1.5	1.9
Supplier Origin	Supplier for the purchase Part	2.5	2.4
Arrival Time	Arrival time of the shipment	-6.5	-4.2

3.3. Constraint and Limitations

The biggest challenge in supervised learning is that irrelevant input features from real-world data classification to present training data could give inaccurate results. The decision boundary might be overstressed if your training set does not have samples that you want to have in a class." The broader philosophy of data preparation is to discover how to best expose the underlying structure of the problem to the learning algorithms" [6]. Hence, Data preparation and processing is always a challenge. Accuracy undergoes when impossible, unlikely, and incomplete values have been recorded as training data. The first steps of our process of elaborate data cleaning and deciding which variables to preserve for the final modeling. A great lot of the data was categorical, and many of the variables were totally connected with others (For example. Part, Supplier and Buyer Destination Country) meaning that we had to be thorough in ensuring that zero in our final dataset was redundant. Furthermore, the large number of categorical variables also destined that many dummy variables would have to be formed. In addition, many of the columns, mainly involving dates, had up to 80% of the data not present, which had to be accounted for as well. Furthermore, strong outliers occurred as a result of unclean data, which led to lead time records of as little as zero days to multiple years, with either not likely in this scenario. The main data manipulation techniques extensively used in ML algorithm are filtering and ordering rows, renaming and adding columns, and computing summary statistics. However, "there are 8 essential data handling verbs that were used to complete the data manipulations job. Such as 1) filter(); 2) distinct(); 3) arrange(); 4) select(); 5) rename(); 6) mutate();7) summarize() and 8) group_by()" [7].

4. Conclusions

In this paper, we demonstrate that machine learning can accurately predict end-to-end Parts shipments delivery lead times. The proposed methods can reduce the time and cost of overall logistics for the equipment movement. The developed scripts can be implemented in real time or near real time in the Oil and gas industry. Furthermore, if the identified approach is also suitable for different processes, the scope of the analysis can be extended using a similar approach.

5. Recommendation

We recommend further experimental and theoretical studies of the machine learning for lead time prediction in Oil and gas industry. Additional training set for accurate sampling is required to achieve the higher accuracy.

References

[1]. Reference: https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

[2]. Zhong, R., Johnson, R. L., & Chen, Z. (2020, June 1). "Using Machine Learning Methods To Identify Coal Pay Zones from Drilling and Logging-While-Drilling (LWD) Data". Society of Petroleum Engineers. doi:10.2118/198288-PA

[3]. Lukas Lingitz,Viola Gallina,Fazel Ansari,Dávid Gyulai,András Pfeiffer,Wilfried Sihn,László Monostori."Lead time prediction using machine learning algorithms: A case study by a semiconductor manufacturer". Publication: Procedia CIRP-Publisher: Elsevier-Date: 2018

[4]. Cheng, Y., Chen, K., H., S., Zhang, Y., Tao, F. Data and knowledge mining with big data towards

smart production. Journal of Industrial Information Integration 2017;doi:10.1016/j.jii.2017.08.001.

[5]. Alexei Botchkarev." Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology". Interdisciplinary Journal of Information, Knowledge, and Management .[v14] [Sep 2018]

[6]. Jason Brownlee. Data Preparation for Machine Learning. https://machinelearningmastery.com/data-preparation-for-machine-learning/. [Online]. [Edition 1.2]. [Sep 21,2020]

[7]. Mathworks. Mastering Machine Learning: A Step-by-Step Guide with MATLAB. https://www.mathworks.com/content/dam/mathworks/ebook/gated/machine-learning-workflow-ebook.pdf. [Online]. [Edition 1]. [Apr 05, 2020]