The European Journal of Research and Development

journals.orclever.com/ejrnd

OR CLEVER Science & Research Group

Research Article

Diagnosis Prediction of Construction Vehicles and Model Explainability Industry 4.0 Implementation

Esra Akca[1*], Semra Erpolat Taşabat[2*], Tayfun Özçay[3*], Nermin Yalçı[4*]

[1] BorusanCAT R&D, İstanbul, Turkey, (ORCID: 0000-0003-3790-8584), esakca@borusan.com
[2] Mimar Sinan Fine Arts University, Faculty of Science, Departmant of Statistics, İstanbul, Turkey, (ORCID: 0000-0001-6845-8278), semra.erpolat@msgsu.edu.tr
[3] BorusanCAT R&D, Information Technologies Director, İstanbul, Turkey, (ORCID: 0000-0003-0011-7877), tozcay@borusan.com
[4] BorusanCAT R&D, İstanbul, Turkey, (ORCID: 0000-0001-6444-4602), nyalci@borusan.com
[*] Correspondence: esakca@borusan.com

(First received October 06, 2021 and in final form March 28, 2022)

Atıf/Reference: Akca, E. Taşabat Erpolat, S. Ozcay, T. Yalci, N. (2022). Diagnosis Prediction of Construction Vehicles and Model Explainability Industry 4.0 Implementation. The European Journal of Research and Development, 2(1), 5-33.

Abstract

With the developing technology, the production and use of computerized machines and vehicles have gathered momentum and consequently the amount of information obtained has increased gradually. The conversion of continuously increasing and streaming data stacks into logical and useful information becomes more and more important. Various methods and algorithms have been developed for this purpose. This group of methods and algorithms, which are called data mining in the most general terms, have been combined with statistics and turned into methods of more comprehensible and logical solutions.

Today, many devices produced are equipped with electronic circuit elements and software. Thus, the operation and control of these devices becomes convenient and it is possible to detect any failure that may occur in the devices in such a way that it does not stop the flow of the process. This has resulted in significant returns in terms of cost and time.

BorusanCAT, one of the world's leading companies in the production and maintenance of construction vehicles, aims to use electronic circuits in its equipment in the most effective way. Construction vehicles of Caterpillar Inc. sold by BorusanCAT are equipped with sensors used to detect major faults that require replacement. With the help of these sensors, the data related to the operation of the construction vehicles are provided instantly via satellite over GPS and GPRS.

In this study, an anomaly detection model has been developed to provide early fault detection and vehicle maintenance needs by using instant data obtained from Caterpillar Inc. construction machinery (vehicles).

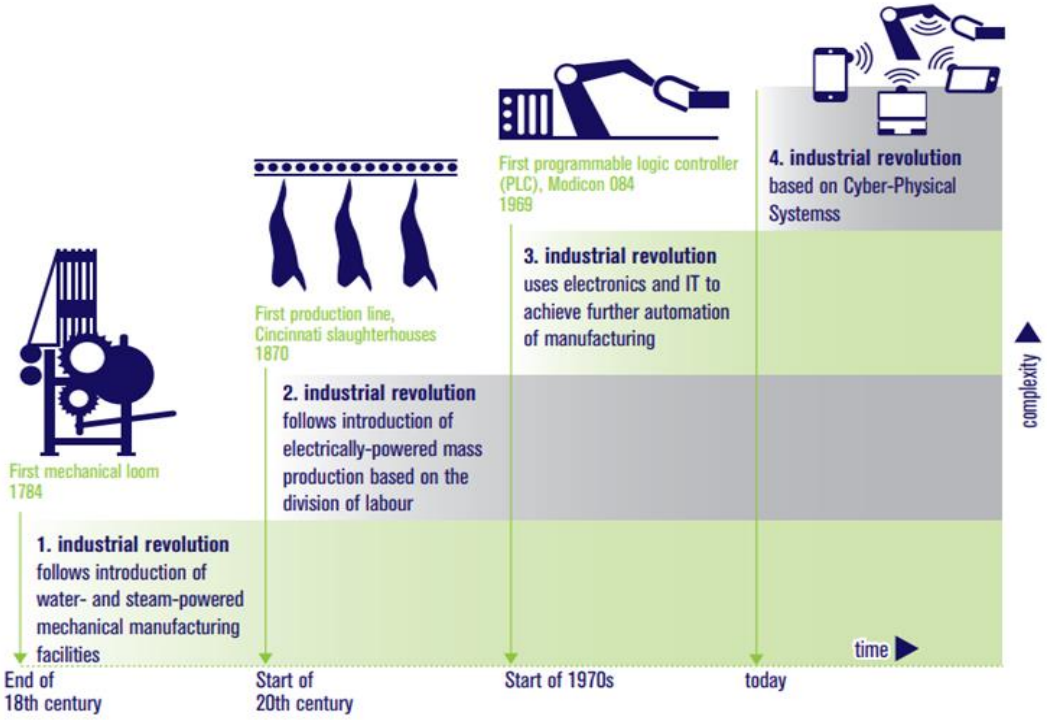
With the Early Warning System (EWS), primarily, the selected sensor data coming from the satellite related to the vehicles is used to predict the failure possibility of the vehicles in a certain time ahead remotely by using the methods of machine learning methods and using the internet of things and cloud technology. After all, prediction data is integrated into business process decision-making with the addition of model explainability.

Keywords: Machine Failure, Machine Learning, Early Warning System, Anomaly Detection, Model Explainability

1. Introduction

Industry 4.0 is emerging as a new vision with developing technology. Industry 4.0 was first used in 2011 at the Hannover Messe in Germany. Its aim is to achieve successful results in cooperation by enabling devices and machines to communicate with each other and with humans. Sustainable production, real-time data monitoring, system intervention before a failure occurs, high efficiency, reduction in production costs and product launch time, and flexibility in production have made Industry 4.0 advantageous.[1]

The development of the industry is briefly summarized in Figure 1.[2] As can be seen from the figure, the development of the industry started in 1784 with the development of the production industry based on steam engines. Then, the development gained momentum in 1870, with the development of electrical energy-based factories and production lines, and in 1969 with electronics and automation-based production, and finally reached its highest level in 2011 with cyber-physical systems (Industry 4.0).



[1] Internet: What is Industry 4.0?, (2018), from www.armolis.com/akademi/endüstri-40

[2] Internet: Recommendations for implementing the strategic initiative INDUSTRIE 4.0, (2013), from http://alvarestech.com/temp/tcn/CyberPhysicalSystems-Industrial4-0.pdf

Figure 1 Industrial Revolution.

In 2013, with the support of the German Federal Ministry of Education and Research, Industry 4.0 study group (Prof. Dr. Henning Kagermann, Prof. Dr. Wolfgang Wahlster, Dr. Johannes Helbig) proposed strategic implementation for the Industry 4.0 initiative. Many academicians and sector representatives participated in the study. This initial study laid the foundations for the ongoing new industrial revolution under different names all over the world. It has arisen from the need for a new and theoretical basis for the research, development of and design of complex industrial systems that can be adapted very dynamically to specific production needs.

In this context, the three main definitions that stand out - the internet of services, the internet of things and cyber-physical systems (CPS)- have been adopted worldwide. The internet of services serves to represent the services and functions as sensitive software components and is offered via the cloud. Development platforms enable the development of web-compatible services. The internet of things is the technological vision for integrating all kinds of objects, devices and persons in a universal digital network. It is intended to help people in the future. The Internet of things and services mainly targets companies and public administrations. Cyber physical systems; communication objects, devices, machines and logistics components that consist of integrated systems. It can communicate over the Internet and using Internet services, form a network with each other and make autonomous decisions with people in a decentralized manner. Cyber-physical systems help a series of tiring, slow, and insecure jobs for humans to be done or supported by machines. It is able to carry out its tasks autonomously by making decisions at the micro level as much as possible.

Contacting with the most of the principles that Industry 4.0 is based upon, this study is one of the most comprehensive Industry 4.0 studies conducted in Turkey. It was discovered by integrating industrial processes and information technologies. The work generally involves the steps of remote sensing, performance monitoring, and execution of tasks in real time. Immediate data flow from over 10,000 construction vehicles is provided by GPS and GPRS by using internet of things technologies across Turkey. The machinery is monitored instantly with more than 1600 different sensor types, tens of them per machine. The size of the generated data stream is more than 2 million data per minute. From this perspective, this study can be considered as a big data project in terms of data amount. However, this large amount of data alone does not mean much. For this

reason, big data is not about how much information is available, but it focuses on what can be done with that information and provides the right approach in decision making processes for saving money, time and sustaining the system. Accordingly, the streaming data is made ready for modelling via various data manipulation methods. The discovery process of the EWS proposed in this study consists of the following steps:

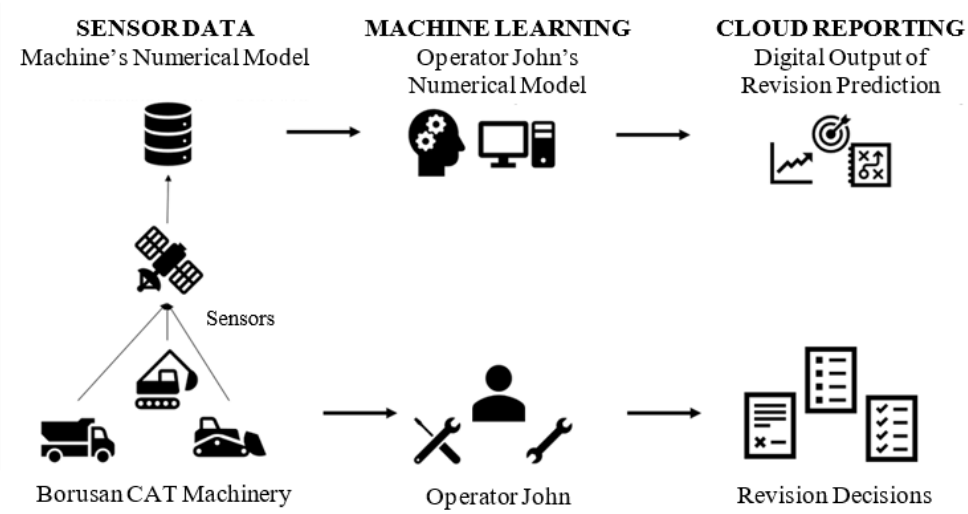


Figure 2 EWS Knowledge Discovery Process

2. Literature Research

There are different studies in this field in the literature. The most noteworthy of these are listed below.

Monitoring of the state of a motor vehicle using machine learning and data mining technology to generate component models that are then used to monitor components, predict failure, etc., such analysis being useful for repair, etc.

Component data analysis involves: provision of a version of a behavior model of the component; prediction of component behavior based on the first model version; collection of component performance data; comparison of the predicted behavior with the collected data; determination of any discrepancies between the two; determination of whether the discrepancy stems from a component failure; and if not modification of the first model version. Independent claims are also made for a component data analysis

system, a method for monitoring component service lives and a diagnosis method for application to motor vehicle components so that their service life can be estimated.[3]

Method and system for condition monitoring of vehicles;

A method and system for an improved vehicle monitoring system in order to provide a cost-effective and scalable system design for industrial application through the use of machine learning and data mining technologies on data acquired from a plurality of vehicles to create models. Frequent acquisition of vehicle sensor and diagnostic data enables comparison with the created models to provide continuing analysis of the vehicle with respect to repair, maintenance and diagnostics. [4]

Service Life predicting method and system for machine and vulnerable component of generating set;

The invention relates to a system for estimating the life of the mechanical and electrical wearing parts for the power set, which comprises a database, a server, a software, a knowledge base, a man-machine interface, and a user terminal system, the life estimation method comprises: counting the history data of the life of the mechanical and electrical wearing parts, setting up a life database of the mechanical and electrical wearing parts, determining the distribution category and parameters of the mechanical and electrical wearing parts, counting the reliability R(t) of the mechanical and electrical wearing parts, counting the reliable life tRO of the mechanical and electrical wearing parts, setting up a knowledge base of the malfunction result of the mechanical and electrical wearing parts, determining the changeover period HRP of the mechanical and electrical wearing parts, estimating the residual life HRL, adjusting the planned maintenance interval to realize the optimized maintenance. The advantages of the invention are that the changeover period of the mechanical and electrical wearing parts can be calculated quantitatively in the using stage of the power set, and the residual life of the parts can be calculated quantitatively, thereby a computer online monitor of the

[3] Achim Bertsche, & Thorsten Engelhardt, (2001). Monitoring of The State of a Motor Vehicle Using Machine Learning, US.

[4] Claude-Nicolas Fiechter, & Mehmet H. Göker, (2001). Method and System For Condition Monitoring of Vehicles, US.

changeover period and the residual life of the mechanical and electrical wearing parts can be realized. [5]

Systems and methods for analyzing equipment failures and maintenance schedules;

A computer implemented method may be used for analyzing equipment failures and maintenance schedules. An equipment maintenance system generates a model of equipment and components of each piece of equipment. In one embodiment, the model is a tree representation. The equipment maintenance system may then determine estimated failure information for each component based on a selected statistical model. The equipment maintenance system may also generate a maintenance schedule based on the determined estimated failure information for each component of the equipment. In one embodiment, the equipment maintenance system displays the equipment maintenance information. [6]

Failure diagnosis method, failure diagnosis apparatus, conveyance device, image forming apparatus, program, and storage medium;

A failure diagnosis method diagnoses a failure occurring in a diagnosis target apparatus including a drive mechanism having a drive member that receives power supply to operate and a power transmission member that transmits drive force of the drive member to another member. The method includes automatically acquiring by a sensor an operation state signal indicating an operation state during the drive mechanism operating for a predetermined period; and analyzing the automatically acquired operation state signal based on a failure probability model, which is obtained by modeling a cause of failure occurring in the diagnosis target apparatus with using probabilities, to execute failure diagnosis with respect to each of constituent members of the drive mechanism. [7]

[5] Shi Jinyuan, Yang Yu, Deng Zhicheng, & Wang Xingping, (2007). Service Life Predicting Method and System for Machine and Vulnerable Component, CN.

[6] Wynn G. Richards, Kurt D. Martin, William L. Lieurance, & Melvin M. Kosler, (2011). Systems and Methods for Analyzing Equipment Failures and Maintenance Schedules, USA.

[7] Kaoru Yasukawa, Kouji Adachi, Kouki Uwatoko, Norikazu Yamada, Eigo Nakagawa, & Tetsuichi Satonaga, (2004). Failure Diagnosis Method, Failure Diagnosis Apparatus, Conveyance Device, İmage Forming Apparatus, Program, and Storage Medium, USA.

Intelligent fluid sensor for machinery diagnostics, prognostics, and control;

A system that facilitates device and/or machinery diagnostics, prognostics and control by way of condition sensing, such as sensing the condition of the device and/or a fluid of the device (e.g., fluid health indicators). The system can employ a plurality of sensors to determine a current state and estimate a future state of the fluid and/or device, as well as providing control of the device, e.g., in order to increase the remaining useful life of the fluid and/or operation of the device. The sensors can communicate wirelessly with each other, with the device, and/or with a central control system that provides, e.g., sensor fusion, prognostics and control integration. In addition, the sensors can be powered locally based upon the physical or chemical properties of the environment. [8]

Industrial process data acquisition and analysis;

A method for optimizing an industrial process data is disclosed. The method includes collecting data from a plurality of sensor elements, wherein each sensor element collects data from a portion of the industrial process and verifying the data collected. The method further includes analyzing the data collected for efficiency and generating at least one recommendation for optimizing the industrial process. The method further includes presenting the at least one recommendation generated to an administrator of the industrial process. [9]

3. Data Abstraction and Comprehension of Data

3.1. Data Acquisition

By storing the signals received from the satellite in the cloud, machine/system data are obtained using cloud computing. At the date of registration, the sensor codes obtained from the built-in sensors and microchips for the respective machine and which transmit signals are stored in the data warehouse in the format given in Table 1.

[8] Frederick M. Discenzo, Dukki Chung, Martin W. Kendig, & Kenneth A. Loparo, (2006). Intelligent Fluid Sensor For Machinery Diagnostics, Prognostics And Control, USA.

[9] Eric M. Rumi, Paul J. Zepf, John Baird, Chris Norris-Lue, Tim Rintjema, & Yvonne Wu, (2005). Industrial Process Data Acquisition and Analysis, USA.

Table 1 Sensor codes and an image of raw data.

Serial Number	Model	Received Time	Date	Mid	Cid	Fmi	Eid	Fault	Service date	Service t-sensor t
AXG01341	980H	15.07.2018 06:15	20180715	30	129	1	96	Engine	20180802	17
AXG01341	980H	5.07.2018 08:59	20180705	30	190	22	96	Transmission	20180802	27
SXG01341	980H	17.07.2018 09:58	20180717	30	155	2	96	0	-	
SXG01341	980H	22.06.2018 05:16	20180622	45	2	1	44	0	-	
RXG01432	315A	14.07.2018 06:15	20180714	20	100	1	22	0	20180713	29
RXG01432	315A	8.07.2018 08:59	20180708	20	33	2	222	Injector	20180713	35
RXG01432	315A	16/07/2018 09:58	20180716	20	20	2	960	Injector	20180713	27
NXG01432	315A	20/06/2018 05:16	20180620	20	10	3	13	0	-	

The explanations about the variables in the data set are as follows:
MID : Signals from integrated microchips on the machine to which sensors are connected.
CID/EID : Sensors used for diagnostic purposes.
FMI : Sensor signal strength.
The variables listed above are considered as independent variables within the scope of the study and the variable containing machine failures collected under the basic codes given in Table 2 is considered as the target variable.

Table 2 Equipment faults and codes

Component Code	Component Name
1000	Engine
3000	Transmission & Drive Line
3030	Transmission
5070	Piston Pump
5459	Swing Drive

3.2. Data Preprocessing (Data Manipulation)

Before the development of the EWS, the pre-processing procedure of the raw data from the sensors is to be carried out.

The details of this procedure, which consists of two steps - data acquisition and recoding of the data - are as follows:

Step 1: Data acquisition

Since it has been indicated that faults are being reported after machines had failed and the period for service records being recorded and component being repaired and registered to the system last a maximum of 30 days, the t-30-day signal data prior to failure notifications have not been taken into consideration.

For non-defective machines, the 30-day data before the model production date have not been taken into account. Since the machine sensors generate a continuous signal for a certain period of time or until they are repaired, the last fault signal generated from each sensor in the range of t-30 and t-60 is accepted. In order for a machine to be considered faulty, it must generate a signal within the range of t-30 and t -60 from the date of the failure on. Figure 3 shows the process of data acquisition.

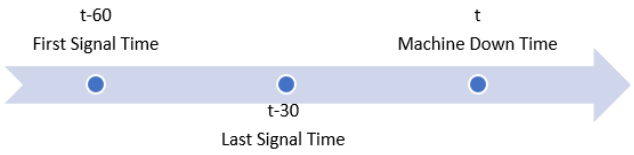


Figure 3 Data acquisition process.

Step 2: Recoding of data

The signal data (CID, FMI, EID) that come under the three main categories are rearranged so that each sensor is a dummy variable.

If the sensor number is written to the data warehouse at the relevant date, the dummy variable of that sensor takes the value 0 if it is not written.

In Tables 3 (a) and (b), the conversion stage of the CID variable is handled. Accordingly, the codes 129, 190, 155 of the CID variable in (a) were converted into dummy variables in (b). The dummy variable has a value of 1 if the sensor number is entered in the data

warehouse and a value of 0 if it is not. The conversion procedures of FMI and EID variables were performed in a similar way.

Table 3 An image from the conversion stages of the CID variable.

Serial Number	Component	CID	Date
RXG01432	Engine	129	1/8/2018
NXG01432	-	190	2/8/2018
NXG01432	-	190	3/8/2018
NXG01432	-	155	3/8/2018

Serial Number	Component	CID-129	CID-190	CID-155
RXG01432	Engine	1	0	0
NXG01432	-	0	1	1

4. Methodology

In the following sub-sections, the models included in the content of the EWS developed within the scope of the study will be presented. Then, in order to make comparisons of all developed and used models, preferred model comparison criteria and the evaluation criteria used to measure the success of the model are to be given. Finally, the parameter and variable selection will be explained in detail.

4.1. Models in Use

Supervised machine learning enables computers to learn from labelled training data without having been explicitly programmed. The tasks covered by this supervised learning are regression and classification. Accordingly, both non-linear decision tree techniques such as gradient boosting, C5.0 and linear machine learning algorithms such as logistic regression will be used in the modelling phase.

The application is based on the comparison of these techniques using multiple classification techniques on multiple dependent variables. For this reason, logistic regression analysis, gradient boosting and C5.0 decision rule derivation algorithm, which represent the three important components of data mining - statistics, machine learning and database technologies - were determined as classification techniques to be used in practice.

4.1.1. C5.0 Algorithm

Number of data and data quality play an important role in the success of the implementations. In line with the need for up-to-date and fast decision making, the opinion gains importance which suggests that the most suitable model is going to be the C5.0 algorithm. The C5.0 algorithm uses the boosting algorithm to increase accuracy, particularly in large data sets. Therefore the C5.0 algorithm is also known as boosting trees. Among its biggest advantages: the algorithm being fast and that it uses memory efficiently. The C5.0 algorithm uses the concept of information gain and entropy reduction to optimally separate nodes. For the variable, k x probabilities are named as respectively.. $X \quad p_1, p_2 \cdots p_k \quad X$ The entropy for the ... variable or quality is given in the following equation:[10]

$$Entropy = H(X) = -\sum_{j=1}^{k} p_j \log_2(p_j) \tag{1}$$

It is assumed that the target quality is subdivided into $T_1, T_2, ..., T_k$ of T sub-clusters depending on the quality of the training set. The weighted average of the information required to determine the class of each T can be calculated as the weighted total number of entropies. The weighted average of the information is given in the following equation.

$$H_s(T) = \sum_{i=1}^{k} p_i H_s(T_i) \tag{2}$$

The information gain is then calculated to perform the separation. Thus, the C5.0 algorithm performs optimal separation by determining the separation criterion with the largest information gain at each decision node. Information gain is given in the following equation.

$$\text{Information Gain } (S) = H(T) - H_S(T) \tag{3}$$

4.1.2. Gradient Boosting

Gradient Boosting is a prediction algorithm that can be used for regression and classification models. It is generally based on the re-estimation of model errors with sequential models and weak estimators, and the reduction of variance as a result of the

[10] Quinlan, J.R. (1996) "Improved use of continuous attributes in C4.5" Journal of Artificial Intelligence Research, 4:77-90.

aggregation of the resulting models with a certain weight.[11] The fact that sequential models are used in the system developed within the scope of the study and that the estimators are quite small, are the reasons for the preference of this algorithm. In the following formula flow, gradient boosting algorithm generalized in decision trees is being demonstrated.[12] It is a machine learning technique for gradient boosting, regression and classification problems. This, combined with weak prediction models, typically creates a model of decision trees. The purpose of any supervised learning algorithm is to identify and minimize a loss function. How mathematics works for Gradient Boosting algorithm is shown in formula (1), formula (2) and formula (3).13 Assume that you have a square error (MSE) defined as follows:

$$Loss = MSE = \sum (y_i - y_i^p)^2 \quad (1)$$

Where y_i = ith target value, y_i^p = ith prediction, $L(y_{i,} y_i^p)$ is Loss function;

Estimates are required so that the loss function (MSE) is minimum. By using gradient descent and updating estimates according to the learning rate, MSE's minimum values can be found.

$$y_i^p = y_i^p + \alpha * \delta \sum_{i=1}^{n} (y_i - y_i^p)^2 / \delta y_i^p \quad (2)$$

Which becomes, $y_i^p = y_i^p - \alpha * 2 * \sum_{i=1}^{n} (y_i - y_i^p)$ (3)

Basically, the estimates need to be updated so that the sum of the residues is close to 0 (or minimum) and that the estimated values are close enough to the actual values.

4.1.3. Logistic Regression

Logistic regression is the analysis method used in linear classification problems in cases where the dependent variable is categorical and in the presence of independent variables determining the outcome. There are only two possible cases. The developed

[11] Tianqi CHEN, (10.2014). Introduction to Boosted Trees, University of Washington

[12] J.H. Friedman, (02.1999). Greedy Function Approximation a Gradient Boosting Machine

[13] Internet: Gradient Boosting, (2018), from https://devhunteryz.wordpress.com/2018/07/11/gradyan-arttirmagradient-boosting/

model is used to explain the relationship between dependent and independent variables. If the data is categorical (nominal/ordinal), non-parametric statistical methods are being used. Since logistic regression analysis aims to estimate the value of categorical dependent variable, a "membership" for two or more groups is estimated. Accordingly, it can be said that one of the aims of the analysis is to classify and the other is to investigate the relationships between dependent and independent variables.[14] Logistic regression is a model-building technique used in statistics. binary (dichotomous), triple and multi-category is a method used to determine the cause-effect relationship with independent variables (Özdamar, 2004). In the logistic regression analysis, the assumption that independent variables are suitable for normal distribution and that covariance matrixes are equal for all groups, is not sought after (Kalaycı, 2008). Therefore, instead of estimating the value of the dependent variable in logistic regression, the probability of the dependent variable taking the value of 1 is tried to be estimated. Since the result obtained is a probability value, it can only take a value between 0 and 1 (Alpar, 2013). The logistic regression model formulation is shown in formula (1), formula (2), formula (3), formula (4).[15]

X : The probability of occurrence of is when the data matrix for the independent variable is. $X = x$ $Y = 1$ π The specific form of the logistic regression model is shown in formula (1), formula (2), formula (3), formula (4) as follows (Hosmer and Lemeshow, 2000):

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \quad (1)$$

Logit conversion is applied to the nonlinear logistic regression function given in equation (1). When performing the logit operation, the natural logarithm of the Odds of an event is taken. Odds is the ratio of the probability that an event will occur to the probability that it will not. In other words For odds Y=1 P(Y=1)/(1- P(Y=1) (Hosmer & Lemeshow, 2000; Menard, 2002).

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_{1x} \quad (2)$$

[14] Mertler, C. A., & Vannatta, R. A. (2005). Advanced and multivariate statistical methods: Practical application and interpretation (third edition). United States: Pyrczak Publishing.

[15] Emre YAKUT, & Eray GEMİCİ, (05, 2017). Predicting Stock Return Classification through LR, C5.0, CART and SVM methods, and Comparing the Methods Used: An Application at BIST in Turkey

The logit transformation function $g(x)$ covers most of the desired properties of a linear regression model. $g(x)$, x Logit ya da, bağımsız değişkenine bağlı olarak $-\infty$ ile $+\infty$ değer alabilen lineer bir denklemdir. Logistic regression model can be written as following in Odds of the independent variable.

$$\frac{\pi(x)}{1-\pi(x)} = e^{\beta_0+\beta_1 x} \tag{3}$$

When the natural logarithm of Odds is taken, the model becomes linear.

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \ln e^{\beta_0+\beta_1 x} = \beta_0 + \beta_1 x \tag{4}$$

4.2. Model Comparison and Evaluation Criteria

4.2.1. ROC Curve and AUC Value

In the study, the ROC curve and AUC value have been used as model selection criteria. AUC value is used frequently in binominal dependent variables and it is essentially a statistic based on type 1 and type 2 error rates.[16] This value, which cannot be calculated for a multinominal variable, has been made computable by following a different path within the scope of the study. Namely: Each target category is considered separately and the AUC value could be calculated then by combining the other categories. This process was repeated for each category and the ROC curve and AUC values were calculated for different binominal target sets as many as the number of categories. Then, the average of these calculated AUC values was used as model success criteria. The average AUC used in the study is given in formula (1).

$$Avg\ \text{AUC} = (AUC1 + AUC2 + AUC3 + ... + AUCn)/n \tag{1}$$

4.2.2. F Value

[16] Internet: Classification: ROC Curve and AUC, (2019), from https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

In this study, the F-value calculated by taking the harmonic average of the precision and recall values have been used as another comparison criterion. The F value is as shown in formula (2).

F = 2(Precision*Recall)/(Precision+Recall) (2)

The precision and recall measurements obtained in the formula are shown in (3) and (4), respectively.

Precision $= TP / TNF$ (3)

Recall $= TP / (TP + FN)$ (4)

The explanations of the abbreviations used in the formulas are as follows:

TP : Accurate Fault Classification Number

TNF : Total Number of Faults

FN : Incorrectly Estimated Number of Fail-Free Classes

FP : Incorrectly Estimated Number of Successful Classes

TN : True Fault Prediction

SC : Success Criterion

RECALL: The concept of precision is generally indicated by the letter p and is calculated as the ratio of the correct results in the information brought to the whole information brought.

PRECISION: The concept of recall is generally indicated by the letter r and is calculated as the ratio of the incoming correct results that are brought to those that have to be brought.

In order to calculate TP and TNF values, it is necessary to determine which category the probability values predicted by the model fall into. Since the data contains negative and positive values and the number of data is not balanced, the probability limit value was determined according to the Odds ratio. The odds ratio was calculated as in formula (5).

$$Odds\ Ratio = \frac{TP / FP}{FN / TN} \quad (5)$$

4.2.3. Over-learning and Recall

On the other hand, in order to determine the over-learning amount of the model and to measure the sampling recall, the percentage difference of the learning and test sample was taken. Over-learning is that the model begins to learn observations rather than patterns. In this case, the data set used for learning is learned. However, such models reduce the likelihood of making a successful prediction when faced with new and previously unseen observations. Over-learning was calculated as shown in formula (6).

$$Overfitting = 1 - \frac{SC_{train}}{SC_{test}} \quad (6)$$

4.3. Selection of Variables and Parameters

In boosting algorithms, it is generally not advisable to pre-select variables based on predictive power. The algorithm selects them automatically. Nevertheless, there is an expectation that the variables may produce a minimum amount of loss value or a certain level of variance. For this reason, the following conditions are taken into consideration in determining the variable list to be used within the scope of the study:

1. Loss value analysis was performed but no loss value could be detected.

2. In the case of defective machines; sensor variables are expected to produce signals at least 4% of the number of faults.

Considering these limitations, the number of variables decreased from 1390 to 50.

4.3.1. Parameter Optimization with Cross-Validation

In the study, two different samplings were used to train and test machine learning algorithms. In contrast, cross validation was used for parameter optimization in the training cluster. This developed methodology is shown in Figure 4.

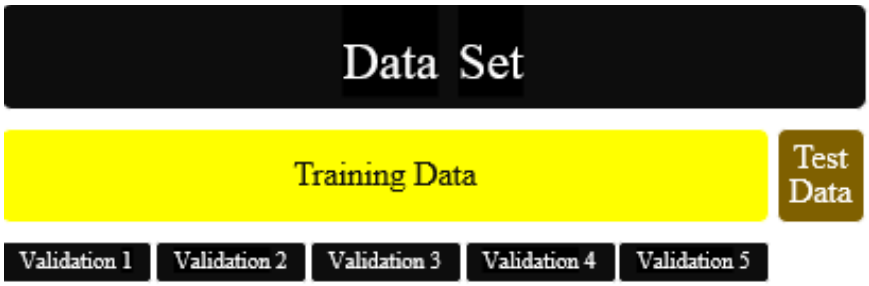


Figure 4 Preparing the data sets

In order to optimize the model parameters, the training set was divided into 5 equal parts with k-fold cross validation, the model was trained with the same parameters by keeping a different part out each time and the external part was estimated. In this way, when the parts that are excluded from the training set and estimated are combined, it covers all observations in the training set and a validation success statistic is calculated.

As a result of repeated iterations with different parameter values, the parameters giving the best validation success statistics were used in the model. Table 4 covers the details of the cross-validation method applied.

Table 4 Cross-Validation

	Training Data Set				
	Validation 1	**Validation 2**	**Validation 3**	**Validation 4**	**Validation 5**
1	Prediction 1	Train	Train	Train	Train
2	Train	Prediction 2	Train	Train	Train
3	Train	Train	Prediction 3	Train	Train
4	Train	Train	Train	Prediction 4	Train
5	Train	Train	Train	Train	Prediction 5

1	Prediction 1
2	Prediction 2
3	Prediction 3
4	Prediction 4
5	Prediction 5

In Table 4, it can be seen that one group of train data, which was divided into 5 equal parts, was excluded in order and five different estimation results were achieved by train data entering the analysis. The combination of estimation results was used to measure success statistics. The formulations of the AUC values used for each cross validation are given below:

AUC1= AUC[C_CODE1000], the code 1000 in the dependent variable is acknowledged as “one”, other codes are acknowledged as “null”

(7)

AUC2= AUC[C_CODE3000], the code 3000 in the dependent variable is acknowledged as "one", other codes are acknowledged as "null"

(8)

AUC3= AUC[C_CODE3030], the code 3030 in the dependent variable is acknowledged as "one", other codes are acknowledged as "null"

(9)

AUC4= AUC[C_CODE5070], the code 5070 in the dependent variable is acknowledged as "one", other codes are acknowledged as "null"

(10)

AUC5= AUC[C_CODE5459], the code 5459 in the dependent variable is acknowledged as "one", other codes are acknowledged as "null"

(11)

The average of these AUC values have formed the average ROC value. The formula (12) has the average ROC value involved.

$$AvgROC = (AUC1 + AUC2 + AUC3 + AUC4 + AUC5)/5 \quad (12)$$

When calculating the average ROC, the AUC values of each category were aggregated and arithmetic means were obtained.

5. Implementing the Models and Results

5.1. Gradient Boosting Model

The results of the model are given in Table 5.

Table 5 Result of Gradient Boosting Model

Iteration	Avg ROC	N.trees	Interaction.depth	N.minobsinnode	Shrinkage
73	82.0%	60	2	15	0.08
1	81.8%	80	3	20	0.05
85	81.7%	110	1	25	0.08
43	81.7%	40	2	15	0.08
4	81.5%	70	2	15	0.08
14	81.5%	100	1	20	0.05
72	81.2%	80	2	10	0.06
84	81.1%	150	1	25	0.1

66	81.0%	110	2	10	0.06
2	71.5%	160	2	45	0.09

The best validation results obtained from 64 different trials with Gradient Boosting algorithm are as follows: Trees: 60, Interaction.depth: 2, Minobsinnode :15, Shrinkage: The average ROC value for 0.08 parameters is 82.0%.

5.2. C5.0 Model

The results of the model are given in Table 6.

Table 6 Result of C5.0 Model

Iteration	AvgROC	Winnow	NoGlobalPruning	EarlyStopping	Trails
1	65.2%	1	1	1	1
43	65.2%	1	1	0	4
4	65.2%	1	1	0	4
14	65.2%	1	1	1	1
42	65.2%	1	1	0	4
47	65.2%	1	1	0	4
16	65.2%	1	1	1	1
44	65.2%	1	1	0	4
41	65.2%	1	1	0	4
64	56.3%	0	0	0	11

The best validation results obtained with the C5.0 algorithm are as follows: winnow:1, noGlobalPruning:1, earlyStopping :1, trails: The average ROC value with 1 is 65.2%.

5.3. Logistic Regression Model

For logistic regression, backward variable selection method was used in each validation sampling. The aim of this analysis is to determine the variables that will provide the optimal cohesion. Variables used in the models have been selected by taking into consideration that these variables are found to be significant in how many validation and that the variables that have been tried in at least 5 models are considered to be stable. The results of the model are given in Table 7.

Table 7 Result of Cross Validation for Logistic Regression Model

Full Validation Set	Validation 1	Validation 2	Validation 3	Validation 4	Validation 5
Intercept	Intercept	Intercept	Intercept	Intercept	Intercept
isold	isold	isold	isold	isold	isold
l2_level_e	l2_level_e	eid119_eid	l2_level_e	l2_level_e	l2_level_e
mid36_mid_e	mid36_mid_e	l2_level_e	mid113_mid	mid113_mid	mid113_mid
mid113_mid	mid113_mid	mid113_mid	mid36_mid_e	mid36_mid_e	mid36_mid_e
c190_8_fmi	eid119_eid	mid36_mid_e	eid284_eid	eid284_eid	c190_8_fmi
eid119_eid	c190_8_fmi	c190_8_fmi	c174_3_fmi	c174_3_fmi	l3_level_e
l3_level_e	l3_level_e	l3_level_e	l1_level	c190_8_fmi	c3547_3_fmi
eid284_eid	mid27_mid	l3_level	mid27_mid	mid27_mid	c174_3_fmi
l3_level	c91_8_fmi	eid284_eid	l3_level	l3_level	l3_level
mid27_mid	mid39_mid_e	mid53_mid	c596_9_fmi	eid265_eid	eid284_eid
c596_9_fmi	mid39_mid	mid39_mid	mid82_mid	c590_9_fmi	mid27_mid
c91_8_fmi	l2_level	mid39_mid_e	mid82_mid_e		mid39_mid_e
mid39_mid_e	l3_level		mid53_mid_e		mid39_mid
mid39_mid	eid284_eid		mid39_mid		
l2_level	mid53_mid		mid39_mid_e		
mid82_mid			l2_level		
			mid263_mid		

6	Logical at all validation and train samplings
5	Logical at all validation samplings
4	Logical at Validation 4 samplings
3	Logical at Validation 3 samplings
2	Logical at Validation 2 samplings
1	Logical at Validation 1 samplings

Table 7 shows the distribution of the significance order of the variables across all validation clusters. The order of significance was scaled within a range of 1 to 6, and the green colour representing 6 predicates significance across all validation clusters whereas the colour representing 1 represents significance in validation 1 samplings.

Table 8 Result of Logistic Regression Model

Variable	Coeficient	Std. Dev	T value	Prob.
İsOld	1.731662	0.147888	11.70931	0.0000

L2_LEVEL_E	1.1510092	0.165821	9.106766	0.0000
MID36_MID_E	-0.654071	0.149915	-4.362935	0.0000
MID113_MID	1.88977	0.312417	6.048875	0.0000
C190_8_FMI	1.051788	0.353036	2.979267	0.0029
MID27_MID	0.949149	0.256952	3.693878	0.0002
MID39_MID_E	-1.028192	0.251712	-4.084792	0.0000
MID39_MID	1.038852	0.246243	4.218811	0.0000
EID284_EID	1.477105	0.400435	3.688748	0.0002
L3_LEVEL	-0.626918	0.216584	-2.894576	0.0038
	Limit Confidence			
LIMIT_2:C(11)	2.184592	0.138888	15.72922	0.0000
LIMIT_3:C(12)	2.592963	0.145968	17.76396	0.0000
LIMIT_4:C(13)	2.757336	0.148853	18.52386	0.0000
LIMIT_5:C(14)	2.819558	0.149993	18.79791	0.0000
LIMIT_6:C(15)	2.868769	0.150892	19.01211	0.0000

Table 8 lists the logistic regression results of the independent variables that are found to be significant in all validation and train clusters that provide the optimal cohesion.

Threshold-limit values determined according to this model are given under "Limit Points". These limit values will help to categorize the variables examined in the sampling.

Table 9 Model Assessment Tools

Model	Parameter	Categories	F-Value (train)	F-Value (Test)	F-Value Overfitting %	AUC (train)	AUC (Test)	AUC Overfitting %
Gradient Boosting	n.Trees: 150	C1000	37.1%	24.3%	53%	93.3%	80.2%	16%
	Interaction.depth : 1	C3000	0.0%	0.0%	0%	98.0%	72.6%	35%
		C3030	0.0%	0.0%	0%	97.9%	26.9%	265%
	Minobsinnode :20	C5070	14.2%	10.3%	38%	99.1%	75.6%	31%
	Shrinkage : 0.08	C5459	12.5%	8.3%	51%	99.5%	65.5%	52%
C5.0	Iteration:62	C1000	27.4%	21.8%	26%	75.7%	72.9%	4%
	Winnow:1	C3000	4.9%	1.7%	188%	71.9%	38.8%	85%
	noGlobalPruning	C3030	6.1%	1.9%	221%	78.5%	64.2%	22%
	earlyStopping no	C5070	15.2%	8.7%	75%	94.9%	30.1%	215%
	trails:10	C5459	11.7%	4.7%	149%	96.0%	66.1%	45%
Logistic Regression	Limit_2:C(11) 2.184592	C1000	21.1%	19.6%	8%	78.2%	75.5%	4%
		C3000	4.3%	2.8%	54%	78.4%	47.7%	65%

Limit_3:C(12)	C3030	5.4%	1.5%	260%	82.8%	66.3%	25%
2.592963	C5070	6.7%	7.5%	-11%	94.4%	81.0%	17%
Limit_4:C(13)	C5459	3.8%	3.1%	23%	90.3%	93.7%	-4%
2.757336							
Limit_5:C(14)							
2.819558							
Limit_6:C(15)							
2.868769							

Continuing to review the results in Table 9, the model success status for categories of the train and test clusters according to the F-value is as follows:

For train data: It can be said that the C1000 category produces best train results via Gradient Boosting algorithm with a ratio of 37.1%. On the other hand, the best test results for the C3000, C3030, C5070 and C5459 categories were obtained with the C5.0 algorithm with ratios of 9%, 6.1%, 15.2% and 11.7% respectively.

Looking at the over-learning values of AUC; it can be said that the Gradient boosting algorithm has over-learned for the component C3030 and over-learned for the components C5070 and C3000 for the C5.0 algorithm.

As a result, if the EWS is run for train data, it can be said that the best model success rates will be achieved when a hybrid structure will be built where according to F value for C1000 category the Gradient boosting algorithm is preferred and for the C3000, C3030, C5070 and C5459 categories the C5.0 algorithm is preferred. In other words, as a result of the implementation of the hybrid structure in question, creating warnings for the construction vehicles to taken to maintenance service before the occurrence any failures is being performed in the most reliable way. Table 10 summarizes the subject situation.

Table 10 Structure of EWS according to the F-Value for train data

Categories	C1000	C3000	C3030	C5070	C5459
Method	Gradient Boosting	C5.0	C5.0	C5.0	C5.0

For test data: It can be said that the C1000 category produces best test results via Gradient Boosting algorithm with a ratio of 24.3%. On the other hand, the best test results for the C3030, C5070 and C5459 categories were obtained with the C5.0 algorithm. The ratios are 1.9%, 8.7% and 4.7% respectively. It can be said that the C3000 category achieved the best test result on the logit regression model with a ratio of 2.8%. Figure 5 aims to explain this situation visually.

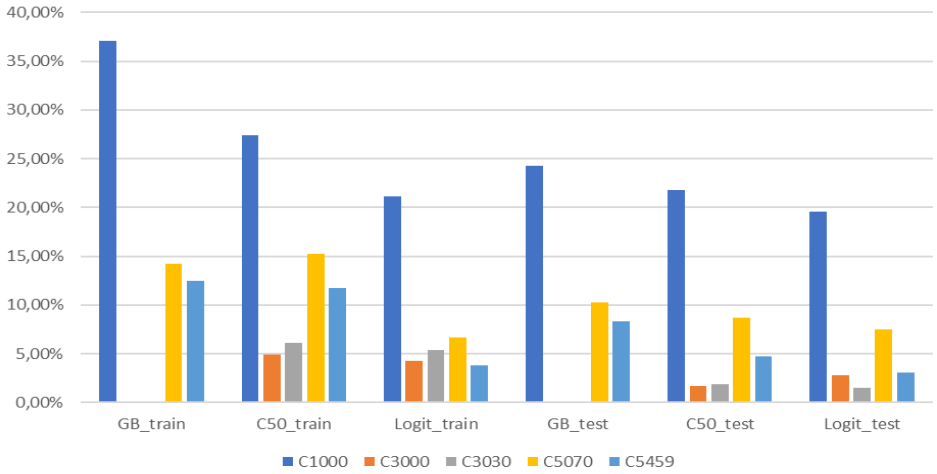


Figure 5 Comparison of methods according to the F-value.

Table 11 shows the architectural structure of the EWS according to the F-value for the test data.

Table 11 Structure of EWS according to the F-Value for test data

Categories	C1000	C3000	C3030	C5070	C5459
Method	Gradient Boosting	Logistic Regression	C5.0	C5.0	C5.0

Continuing to review the results in Table 9, the model success status for each category of the train and test clusters according to the AUC value is as follows:

For train data: The best train results for all categories were 93.3% (C1000), 98.0% (C3000), 97.9% (C3030), 99.1% (C5070) and 99.5% (C5459) respectively. The architecture of the EWS is shown in Table 12.

Table 12 Structure of EWS according to the AUC-Value for train data

Categories	C1000	C3000	C3030	C5070	C5459
Method	Gradient Boosting	Gradient Boosting	Gradient Boosting	Gradient Boosting	Gradient Boosting

For test data: It is clear that the Gradient Boosting algorithm produces the best test results in the C1000 and C3000 categories with ratios of 80.2% and 72.6% respectively. In the C3030, C5070 and C5459 categories, the Logit Regression model gives the best results with ratios of 66.3%, 81.0% and 93.7% respectively. Figure 6 aims to explain the results briefly. The architectural structure of the EWS is shown in Table 13.

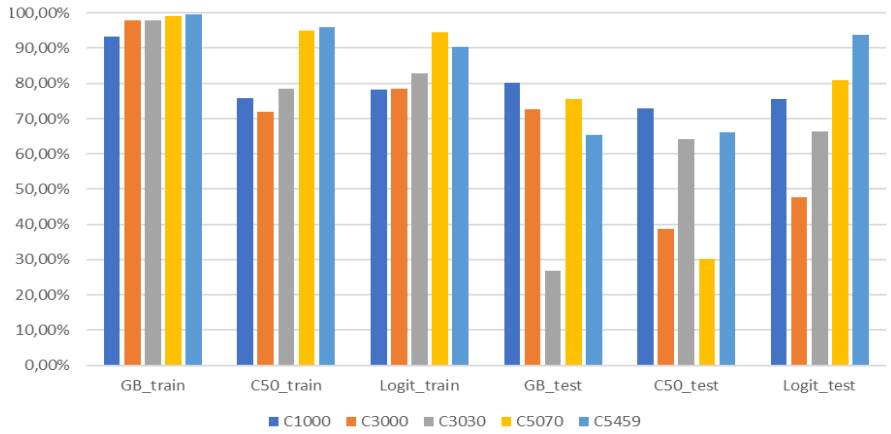


Figure 6 Comparison of methods according to the AUC value.

Table 13 Structure of EWS according to the AUC-Value for test data

Categories	C1000	C3000	C3030	C5070	C5459
Method	Gradient Boosting	Logistic Regression	Logistic Regression	Logistic Regression	Logistic Regression

6. Model Explainability

Explainability and interpretability; It refers to the extent to which the internal mechanics of a machine or deep learning system can be explained in human terms. As AI prediction models continue to make increasingly accurate predictions, there is growing

demand to apply them to mission critical tasks such as credit screening, crime prediction, fire risk assessment, and emergency medicine demand prediction. For such applications, recent complex prediction models are often criticized as black boxes as it is difficult to understand why the prediction model produces a given output. Aim; It is to explain which variables change the output of machine learning algorithms called black boxes. Shapley uses only observed data points. Performs reliable explanations in line with human intuition. Developed by Lundberg and Lee, SHAP (SHApley Additive ExPlanations) is a method for explaining individual predictions. SHAP is based on the game's theoretically optimal Shapley Values. Lundberg et al. They proposed TreeSHAP, a variant of SHAP for tree-based machine learning models such as decision trees, random forests, and gradient supported trees.[17] Explaining the output of machine learning algorithms is a game theory approach. The importance of the values is made with the Shapley values. For example, which factors are more important for an individual to have diabetes. Here, the factors may have had negative and positive effects. The important thing is that it affects.

Property attributes such as Shapley values can be visualized as "forces". Each attribute value is a power that increases or decreases the estimate. Estimation starts from baseline. The basis for Shapley values is the average of all estimates. In the drawing, each Shapley value is an arrow that pushes the estimate up (positive value) or down (negative value). These forces balance each other in the actual estimate of the data sample.

Black box prediction models used in statistics, machine learning and artificial intelligence have been able to make increasingly accurate predictions, but it remains hard to understand those predictions. The purpose of explaining the model in practice is expressed in the Figure 7.

[17] Lundberg, Scott M., Gabriel G. Erion, and Su-In Lee. "Consistent individualized feature attribution for tree ensembles." arXiv preprint arXiv:1802.03888, 2018.

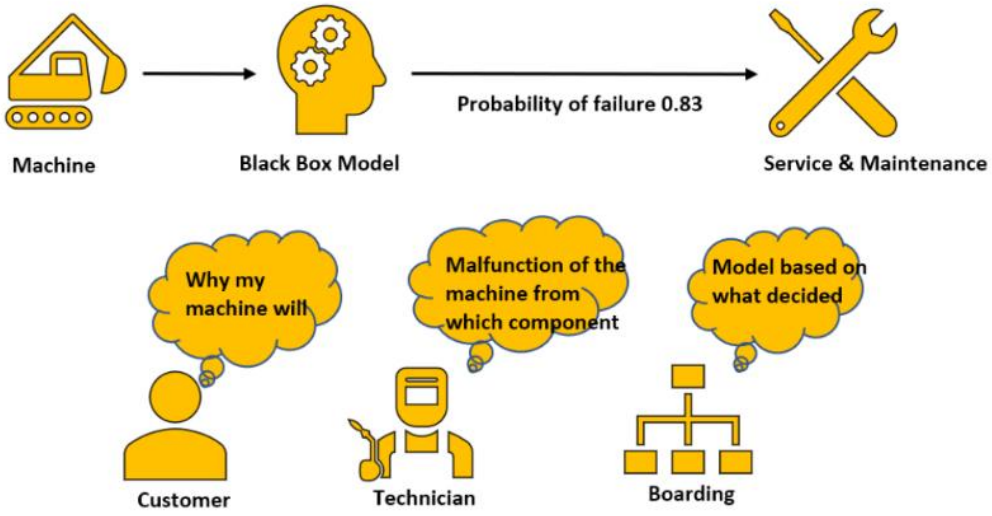


Figure 7 The Explanatory Purpose of the Model in Practice

In the application, besides estimating the probability of failure of the work machines, the SHAP explanation power graphs that will cause this failure are shown. Below we see the result of a sample machine.

Figure 8 SHAP Values to Describe the Predicted Probability of Failure of the Machine

In Figure 8; The baseline – the average predicted probability – is 0.84. The estimated ric of the vehicle is 0.84. Risk increasing effects such as ZMODEL (machine model group) are balanced by reducing effects such as WORKH_CAT (machine working hours).

SHAP has a solid theoretical foundation in game theory. The prediction is fairly distributed among the feature values. It also helps to unify the field of interpretable machine learning.SHAP has a fast implementation for tree-based models. SHAP is integrated into the tree boosting frameworks xgboost and LightGBM. In this way, the explainability of the application is provided.

7. Conclusion

Data mining covers important points about the accurate, effective, purposeful use of the data and what to do with the information, rather than the size of the data. In this study, where machine learning is used effectively, it can be seen that all 3 methods can achieve better results for different categories when different success criteria are being considered. It is important whether the criterion selected serves the performance logic to be measured. For example, the AUC value attaches importance to the whole distribution as it deals with the ordering of the wrong estimates, while the F value focuses only on false estimates above or below the specified limit value, as it is focused on the decision matrix.

In order to choose the accurate model, it is necessary to make the train set results the basis before moving forward. When the AUC value is considered as the criterion of success for the train set results, it is clear that the best algorithm is Gradient Boosting for all categories. Gradient boosting produces the best result within the C1000 category for the F-Value while C5.0 does that in the other categories.

In the test results, when the AUC criterion is taken into consideration, it is clear that the Logit and Gradient Boosting algorithms can be used together to produce the best results in certain categories. Taking F criteria into consideration, it is clear that using all three model approaches together for different categories produces the best result.

In the model used, the data from the sensors in the construction vehicles are trained and learn the patterns obtained from machines under maintenance to predict the vehicles that need possible maintenance in the future.

In summary, it is certain that the use of different algorithm approaches together will be beneficial in optimizing the results and reducing over-learning. A combination is possible for the same categories, such as the use of different algorithm results for different categories. In general, the main purpose of machine learning applications is learning patterns from data and using these patterns to create value. For this purpose, by revealing valuable, usable information from the big data; the decision support system has been evaluated.

References

[1] Internet: What is Industry 4.0?, (2018), from www.armolis.com/akademi/endüstri-40

[2] Internet: Recommendations for implementing the strategic initiative INDUSTRIE 4.0, (2013), from http://alvarestech.com/temp/tcn/CyberPhysicalSystems-Industrial4-0.pdf

[3] Achim Bertsche, & Thorsten Engelhardt, (2001). Monitoring of The State of a Motor Vehicle Using Machine Learning, US.

[4] Claude-Nicolas Fiechter, & Mehmet H. Göker, (2001). Method and System For Condition Monitoring of Vehicles, US.

[5] Shi Jinyuan, Yang Yu, Deng Zhicheng, & Wang Xingping, (2007). Service Life Predicting Method and System for Machine and Vulnerable Component, CN.

[6] Richards, W. G., Martin, K. D., Lieurance, W. L., & Kosler, M. M. (2013). U.S. Patent No. 8,494,826. Washington, DC: U.S. Patent and Trademark Office.

[7] Yasukawa, K., Adachi, K., Uwatoko, K., Yamada, N., Nakagawa, E., & Satonaga, T. (2012). U.S. Patent No. 8,132,049. Washington, DC: U.S. Patent and Trademark Office.

[8] Discenzo, F. M., Chung, D., Kendig, M. W., & Loparo, K. A. (2009). U.S. Patent No. 7,581,434. Washington, DC: U.S. Patent and Trademark Office.

[9] Rumi, E. M., Zepf, P. J., Baird, J., Norris-Lue, C., Rintjema, T., & Wu, Y. (2007). U.S. Patent No. 7,218,974. Washington, DC: U.S. Patent and Trademark Office.

[10] Quinlan, J. R. (1996). Improved use of continuous attributes in C4. 5. Journal of artificial intelligence research, 4, 77-90.

[11] Chen, T. (2014). Introduction to boosted trees. University of Washington Computer Science, 22(115), 14-40.

[12] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

[13] Internet: Gradient Boosting, (2018), from https://devhunteryz.wordpress.com/2018/07/11/gradyan-arttirmagradient-boosting/

[14] Mertler, C. A., & Vannatta, R. A. (2005). Advanced and multivariate statistical methods: Practical application and interpretation (third edition). United States: Pyrczak Publishing.

[15] Emre YAKUT, & Eray GEMİCİ, (05, 2017). Predicting Stock Return Classification through LR, C5.0, CART and SVM methods, and Comparing the Methods Used: An Application at BIST in Turkey

[16] Internet: Classification: ROC Curve and AUC, (2019), from https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

[17] Lundberg, Scott M., Gabriel G. Erion, and Su-In Lee. "Consistent individualized feature attribution for tree ensembles." arXiv preprint arXiv:1802.03888, 2018.