International Journal of Advanced Smart Convergence Vol.12 No.2 47-55 (2023)
http://dx.doi.org/10.7236/IJASC.2023.12.2.47

IJASC 23-2-5

A Study on the Application of Measurement Data Using Machine Learning Regression Models

Yun-Seok Seo*, Young-Gon Kim**

** Ph. D. Candidate, Department of Computer Engineering, Tech University, Korea*
ysseo@tukorea.ac.kr

***Professor, Department of Computer Engineering, Tech University, Korea*
ykkim@tukorea.ac.kr

Abstract

The automotive industry is undergoing a paradigm shift due to the convergence of IT and rapid digital transformation. Various components, including embedded structures and systems with complex architectures that incorporate IC semiconductors, are being integrated and modularized. As a result, there has been a significant increase in vehicle defects, raising expectations for the quality of automotive parts. As more and more data is being accumulated, there is an active effort to go beyond traditional reliability analysis methods and apply machine learning models based on the accumulated big data. However, there are still not many cases where machine learning is used in product development to identify factors of defects in performance and durability of products and incorporate feedback into the design to improve product quality.

In this paper, we applied a prediction algorithm to the defects of automotive door devices equipped with automatic responsive sensors, which are commonly installed in recent electric and hydrogen vehicles. To do so, we selected test items, built a measurement emulation system for data acquisition, and conducted comparative evaluations by applying different machine learning algorithms to the measured data. The results in terms of R^2 score were as follows: Ordinary multiple regression 0.96, Ridge regression 0.95, Lasso regression 0.89, Elastic regression 0.91.

Keywords: *Machine learning, Regression, Reliability, Automotive*

1. Introduction

In recent years, the automotive industry has been rapidly transforming from a mechanical engineering sector to an electric and electronic devices industry. AI-equipped vehicles are becoming established as new IT devices. Currently, all automotive components in development are keeping pace with the structural changes in this industry. They are either creating new product categories or expanding existing ones to add new features and differentiate themselves from existing products.

However, as automobiles become more advanced, the components required are transitioning from analog

Manuscript Received: March. 19, 2023 / Revised: March. 23, 2023 / Accepted: March. 25, 2023
Corresponding Author: ykkim@tukorea.ac.kr
Tel: +82-31-8041-0530, Fax: +82-31-8041-0539
Professor, Department of Computer Engineering, Tech University, Korea

mechanisms to digital mechanisms. This shift towards digitalization brings about increased complexity and integration in the components, which can lead to a higher frequency of unexpected failures. As shown in Figure 1, according to the data analysis conducted in Seoul over the past five years, it can be observed that vehicle and component defects account for a significant portion compared to simple car failures.

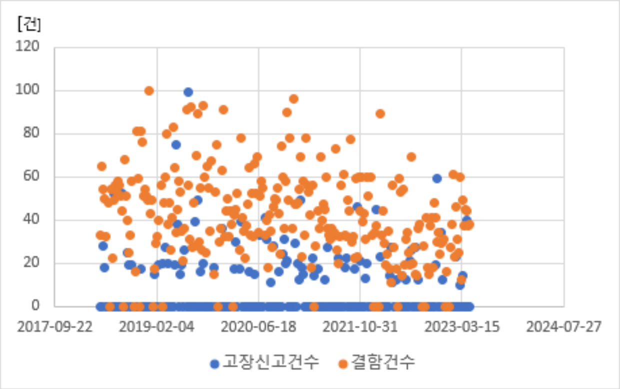

Figure 1. Vehicle breakdown and defect in Seoul comparison chart -For the past five years (2018 ~ 2023) [data source: Google big data]

As a result, the demand for high-quality automotive components is increasing, and the reliability of vehicles is being treated as a crucial aspect[1]. Traditional automotive manufacturers have made various efforts using statistical techniques for defect analysis to ensure product reliability. Currently, with the advancement of research and development in analysis algorithms such as machine learning and deep learning, which utilize big data, there is an opportunity to enhance and complement traditional analysis methods, yielding improved and more accurate results. In machine learning, it is essential to select appropriate test items, conduct suitable tests, and utilize the relevant algorithms to identify the factors of defects (independent variables) in the measured diverse data, ultimately pointing the root causes of anomalies. To enhance the product reliability of touch-sensitive door module components, which are commonly used in electric vehicles and other vehicles, we are interested in researching and evaluating the key factors that influence defect occurrences. This research aims to improve the overall reliability of automobiles. To acquire the necessary data, we designed a test apparatus capable of evaluating reliability using open-source hardware. The purpose was to create a simulation and measurement system that replicates real-time conditions, implementing operating conditions and durability testing environments similar to those experienced by actual vehicles. The developed simulation and measurement system for the door module incorporates electrical and mechanical mechanisms to enable the door module to undergo opening and closing operations, as well as measurement emulation, under actual operational or even harsher conditions. In this paper, the measured data from the sensors was encoded, processed, and applied to an analysis model using machine learning algorithms. Training and prediction were performed using the model, and evaluations were conducted to compare and analyze different machine learning models.

2. Previous Research and principal Background

2.1 Machine learning algorithms

Machine learning is an algorithmic approach that utilizes data to analyze interrelationships, identify patterns, learn rules, and perform predictive analysis. Machine learning algorithms can handle structured data as well as unstructured data, such as internet search queries and social media posts. These algorithms process and transform data into a numerical format that computers can understand. They analyze correlations in the data and use mathematical solutions to discover patterns and select appropriate models. These models are then applied, trained, and validated based on their performance, using the results of validation to utilize the models for problem-solving[2]. Machine learning algorithms can be categorized into supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. Among these, supervised learning algorithms are commonly used for prediction and classification tasks. Some of the key supervised learning algorithms for prediction and classification include k-Nearest Neighbors (kNN), Linear Regression, Logistic Regression, Support Vector Machines (SVM), Decision Trees, Random Forests, and Neural Networks. On the other hand, unsupervised learning algorithms include Clustering, Visualization, Dimensionality reduction, and Association rule learning, among others[3].

In regression analysis, the relationship between the dependent variable (the target variable to be predicted) and the independent variables (the input features used by the model for prediction) is represented by a linear function. The goal is to find the best-fitting line or curve that represents this relationship and can be used to make predictions. The linear function is typically expressed as a linear equation, where the coefficients of the independent variables determine the slope and intercept of the line or curve[4].

$$\hat{y} = \beta_0 + \beta_i x_i \cdots + \varepsilon_i \tag{1}$$

$$[i = 1, 2, 3, \cdots, n]$$

β0 : Intercept of the regression equation, βi : slope(weight), εi : residual
β0, βi : parameters to be learned by the model (linear coefficients)

In this paper, regression analysis algorithms were applied due to the nature of the continuous time-series data. To execute machine learning algorithms, libraries such as Python's scikit-learn (sklearn) were used to preprocess and analyze the data. Additionally, visualization techniques were employed to aid in the analysis process.

3. Propose method and discussion

3.1 System Design and Configuration

In this study, we developed a measurement emulation system to obtain measurement data for a touch-sensitive vehicle door handle device – Retractable Door Outside Handle(RDOH) that is widely used in recent domestic and international electric vehicles. The overall measurement system is shown in Figure 2. To consider the flexibility of the system, we utilized open-source hardware MCU boards such as Raspberry Pi and Arduino.

Additionally, to increase the accuracy of the data, we equipped a high-resolution DAC board with a higher number of bits and resolution. The selection of sensors(Loadcell, Thermocouple, Humidity, Hole effect, etc) was also based on their resolution considerations.

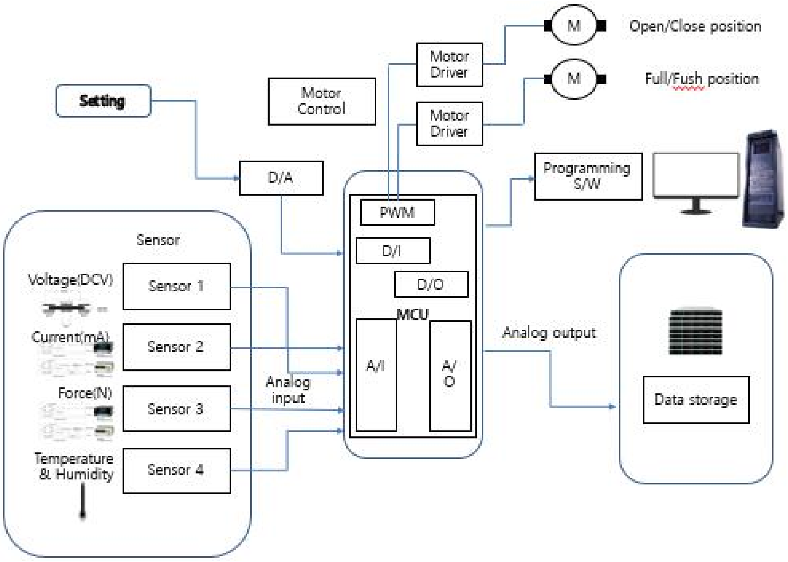

Figure 2. Machine learning measurement system

The acquired measurement data is stored in CSV format and converted into usable data for training machine learning models. It is then split into training data and testing data, which are commonly referred to as training data and test data, respectively.

3.2 Machine Learning Process

The steps of performing machine learning involve analyzing and transforming the characteristics of the acquired data to apply it to a model. The process includes data preprocessing, where the data is cleaned and transformed, and the data is split into training data and test data for model application and evaluation, respectively. The machine learning process involves selecting an appropriate machine learning algorithm model using the training data and performing the learning process. After training the model, predictions are made using the trained model, and evaluation is conducted. The machine learning process follows a sequential order as depicted in Figure 3.

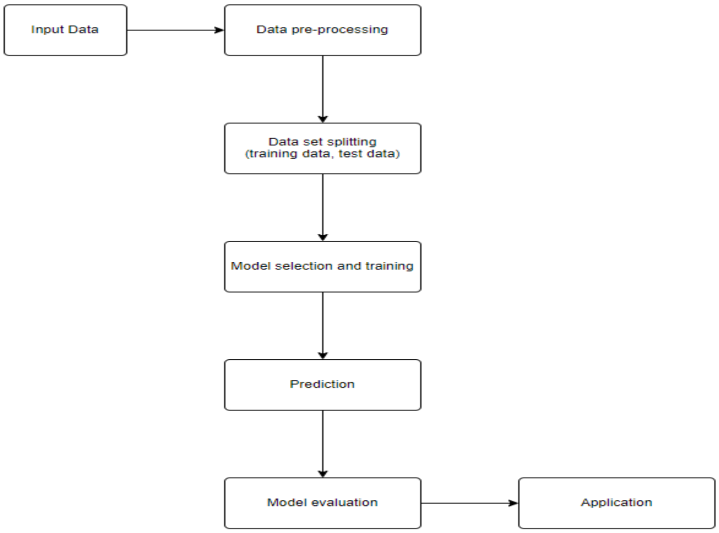
Input Data
Data pre-processing
Data set splitting (training data, test data)
Model selection and training
Prediction
Model evaluation
Application

Figure 3. Procedure for performing machine learning

3.3 Data pre-processing

Data preprocessing is a process that is performed before data analysis. It involves improving the quality of the data, extracting features from the data, and converting the data into a format that is suitable for modeling. If the quality of the data is not good, the performance of the model may be degraded. If features are not extracted from the data, the model may not be able to understand the data or process the data.

In order to analyze the acquired data, it needs to be transformed into a usable form. First, the basic information of the input data is examined to understand the types of data variables. The test items in the collected data that have a pass/fail result value have an object variable value. In a data frame, the "object" type usually represents strings, so they need to be converted to numerical form for analysis. Additionally, data that appears as "object" with attached units should be converted to numerical form. Furthermore, it is important to check for missing values in the data. Missing values are commonly represented as NaN or Null. If such missing values exist, they can make model analysis challenging, so they need to be removed using Python functions.

Since there are unordered categorical data in the features, they need to be converted to numerical form using techniques such as one-hot encoding or label encoding. Additionally, irrelevant items that are not related to the prediction target should be removed. As the units and scales of each feature may vary, it may be necessary to scale them to the same range if needed. The common approach to scaling variables is to normalize the data, which involves subtracting the mean from each variable and dividing it by the standard deviation. This process is known as standardization or z-score normalization.

3.4 Training and Validation data splitting

To train a machine learning model, it is necessary to divide the acquired data into training data and test data. The training data is used for model training, while the test data is used for prediction. The train_test_split() function from the scikit-learn's model_selection module is commonly used to split the data. Typically, a ratio of (70 ~ 80) : (20 ~ 30) is used, where the larger portion is allocated for training data and the smaller portion for test data.

3.5 Model selection and Training

The characteristics of the collected data indicate that it is a time series data type with continuity, and there is a linear relationship between the features and the labels. Therefore, a regression model was selected among machine learning algorithms. The representative linear regression and n-regularized linear models, such as Ridge regression, Lasso regression, and Elastic Net regression, were chosen. Elastic Net regression combines both Ridge regression and Lasso regression regularization, allowing for adjustment of the ratio between the two, which can improve prediction performance.

$$J(\theta) = MSE(\theta) + ra\sum_{i=1}^{n}|\Theta_i| + \frac{1-r}{2}a\sum_{i=1}^{n}\theta_i^2 \tag{2}$$

J(θ) : Cost function of Elastic Net regression
MSE(θ) : Mean Squared Error
When the mixing ratio (r) is 0, it is equivalent to Ridge regression, and when r is 1, it is equivalent to Lasso regression. The mixing ratio adjusts the combination between 0 and 1

The mixing ratio is the ratio of the regularization terms of ridge regression and lasso regression. The closer the mixing ratio is to 0, the closer it is to ridge regression, and the closer it is to 1, the closer it is to lasso regression. By adjusting the mixing ratio, we can prevent overfitting and underfitting of the model.

3.6 Prediction and Evaluation

By applying the trained model to the validation data, we can predict the target values. To evaluate the performance, various evaluation metrics such as MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), R2, etc., are commonly used. It is possible to adjust the parameters to improve the prediction performance.

$$MAE = \frac{1}{N}\sum_{i=1}^{n}\left| Y_i - \hat{Y}_i \right| \tag{3}$$

$\hat{Y}_i$: Predicted value Y_i : actual value
N : Number of observations/rows

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{4}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{5}$$

4. Simulation

The drive and measurement emulation system developed to verify the performance of a capacitive vehicle door device, which is a car component, is depicted in Figure 4.

The data for the product was measured and collected in the system, and machine learning algorithms were applied to the collected data for training.

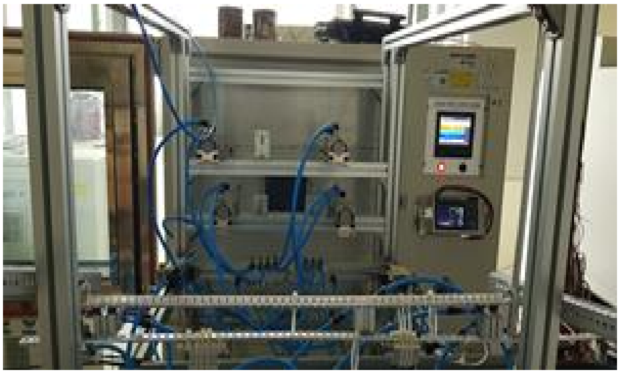

Figure 4. RDOH Measuring System

To conduct technical evaluation, specific measurement criteria were defined, and the types of acquired data are listed in Table 1. The selected criteria for the durability test include the time, force, and noise involved in opening and closing the door handle, as well as environmental conditions such as temperature and humidity.

The handle's impact during the test was also considered as a measurement item.

Table 1. Measured acquisition data Item

Item	data
operating time	time(ms), current(mA), voltage(V)
operating force	Force(N), Distance(mm)
Noise	Noise(dB)
high pressure car washing	Pass/Failure
Temp. & Humidity	Temperature(℃), Humidity(%RH)
in vitro dissolution	Pass/Failure
freeze-thaw	Pass/Failure
internal shock	Pass/Failure
Impact resistance	Pass/Failure
destructive strength	Force(N), Distance(mm)

The measured data was adjusted data preprocessing and analyzed in Python to perform data correlation. The following figure 5 illustrates a scatter plot using a visualization library.

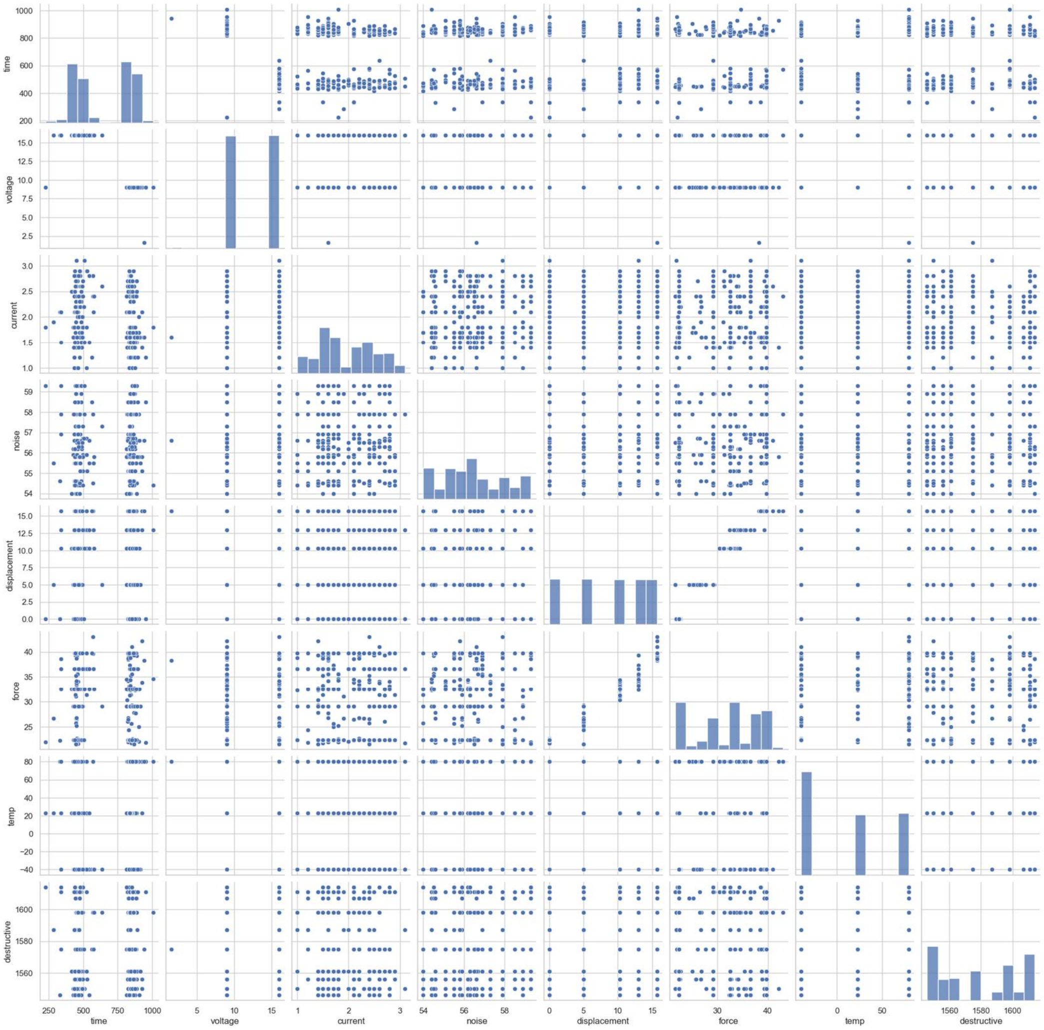

Figure 5. Scatter plot for data correlation.

The object-type variables were converted to numerical values, and due to the varying units of different variables, scaling was performed to make the data consistent. Additionally, through analysis, variables with weak correlations to the target variable were removed. Measured Acquisition data items Are listed Table 2.

Table 2. Measured acquisition data Item

scaler	description
Min-Max Scaler	adjusted to have a minimum value of 0 and a maximum value of 1
Standard Scaler	adjusted to have a mean of 0 and a standard deviation of 1
Robust Scaler	adjusted using the median and the interquartile range (IQR)

We applied various types of scaling and found that the Robust Scaler provided the most suitable results after scaling. After splitting the data into training and validation sets, machine learning algorithms were applied for training. In the case of regularized regression models, the default value for the alpha parameter was used during prediction evaluation. The results showed that the mean squared error (MSE) values were fairly similar across all models. However, the R^2 score values differed slightly. For the ordinary multiple regression model, the R^2 score was 0.96, while for Ridge regression, it was 0.95. In the case of Lasso regression, the R^2 score was 0.89, and for ElasticNet regression, it was 0.91.

R^2 score is an indicator that measures the goodness of fit of a regression model. It has a value between 0 and 1, and the closer it is to 1, the better the fit of the model. R^2 score indicates how well the model explains the data and how much noise there is in the data.

Table 3. Evaluation results

Algorithm	R^2 score
Ordinary multiple regression model	0.96
Ridge regression model	0.95
Lasso regression model	0.89
ElasticNet regression model	0.91

5. Conclusion

With the advancement of artificial intelligence technology, the application of machine learning and deep learning is gradually expanding in various fields. While various analysis techniques have been used in the automotive industry to address issues such as product defects, the use of machine learning algorithms in this field is still relatively limited. In this paper, an open-source hardware and IoT-based data measurement system was developed to assess the performance of the capacitive door latch device module installed in electric vehicles. Experimental data was collected using this system, and since the collected data exhibited a time-series continuity, regression-based machine learning algorithms were applied for analysis. To ensure smooth execution, the data was refined through numerical transformation and scaling operations. Irrelevant variables were removed based on correlation analysis. To determine the optimal algorithm for prediction, we utilized libraries such as scikit-learn in Python to train models and conduct prediction tests on the developed algorithms.

The evaluation results were compared among different machine learning algorithms. Ordinary multiple regression achieved the highest R2 score of 0.96, indicating its significant predictive capability as the score approaches 1. However, in the case of regularized regression models, the numeric values can be increased by adjusting the alpha value. Since the data obtained from simulation equipment has limitations, future research plans to perform comparative evaluations through the accumulation of real-time data and the application of various value adjustments, such as reinforcement learning.

References

[1] Song, M., Kim, K., & Ahn, S., "Changes in the Domestic Automotive Industry Structure: A Text Analysis Approach", Research Report(KIET, Korea Institute for Industrial Economics and Trade), Vol. 2021-10, pp.1-94 2021, DOI:https://doi.org/10.38094/jastt1457

[2] D. Maulud and A. M. Abdulazeez, "A Review on Linear Regression Comprehensive in Machine Learning", JASTT, Vol. 1, No. 4, pp. 140-147, Dec. 2020. DOI:https://doi.org/10.38094/jastt1457

[3] Yong-hee, Han, "Prediction Model of CNC Processing Defects Using Machine Learing", Journal of The Korea Convergence Society, Vol. 13. No. 2, pp. 249-255, 2022, DOI : https://doi.org/10.15207/JKCS.2022.13.02.249

[4] S.-J., Lee, Y.-T., Kim, S.-y., Kim, "Comparison of Customer Satisfaction Indices Using Different Methods of Weight Calculation", The Journal of Digital Policy & Management, Vol. 11, No.12, pp. 201-211, Dec, 2013, DOI:http://dx.doi.org/10.14400/JDPM.2013.11.12.201

[5] Kim, Y.-I., Lee, K.-H., Park, S.-H. (2023) Application and Evaluation of Machine Learning Techniques for Real-time Short-term Prediction of Air Pollutants, Journal of Korean Society for Atmospheric Environment, Vol.39, No.1, 107-127 , DOI:https://doi.org/10.5572/KOSAE.2023.39.1.107

[6] Joong-Soo Lim, "A Design of Small Size Sensor Data Acquisition and Transmission System",Journal of Convergence for Information Technology, Vol. 9. No. 1, pp. 136-141, 2019, DOI:https://doi.org/10.22156/CS4SMB.2019.9.1.136

[7] H. Jie and G. Zheng, "Calibration of Torque Error of Permanent Magnet Synchronous Motor Base on Polynomial Linear Regression Model," in IECON 2019-45th Annual Conference of the IEEE Industrial Electronics Society, , pp. 318-323. 2019, DOI: 10.1109/IECON.2019.8927537