*ICIC Express Letters*
Volume **11**, Number **6**, June **2017**
 ISSN 1881-803X
pp. 1127–1132

# COMPARISON OF QUALITY PREDICTION ALGORITHMS IN MANUFACTURING PROCESS

Aekyung Kim, Kyuhyup Oh, Hoonseok Park and Jae-Yoon Jung*

Department of Industrial and Management Systems Engineering
Kyung Hee University
1732, Deogyeong-daero, Giheung-gu, Yongin-si, Gyeonggi-do 17104, Korea
*Corresponding author: jyjung@khu.ac.kr

Received November 2016; accepted February 2017

Abstract. *Production failure is one of the biggest concerns for every manufacturing company. The production failure may cause a variety of quality costs and lead to production delays. Therefore, one of the most important aims toward smart factory is to develop the quality monitoring system that gives information about upcoming failure of production. In this paper, we deal with a prediction problem of defect rates from manufacturing processing conditions. Representative regression algorithms such as linear regression, non-linear regression and tree-based regression are compared to predict the defective rates for production lots. A real-life dataset of die-casting manufacturing process is used to compare the performance of the regression algorithms. The experimental results show that tree-based regression algorithms generally outperform linear and non-linear regression algorithms. The algorithms can be adopted to implement the quality prediction system for smart factory.*
**Keywords:** Smart factory, Manufacturing quality prediction, Regression algorithms, Die-casting

1. **Introduction.** In manufacturing industries, the huge amount of real-time process condition data is collected from sensors throughout their plant and industrial equipment [1]. The production data is gathered from manufacturing information systems such as manufacturing execution systems (MES). These kinds of data usually involve the valuable information that people want. Furthermore, many manufacturing organizations are looking for ways to improve their quality of production by improving defect tracking and improving forecasting abilities [2]. For that reason, the advanced manufacturing such as demand forecasting, production planning and control becomes more important toward the smart factory [3,4]. Specifically, the defect rate is one of the most significant key performance indicators in manufacturing process. Therefore, it is crucial to develop a prediction model based on the manufacturing condition data in manufacturing quality management [5-7].

In this research, we compare several regression algorithms for quality prediction based on manufacturing data analysis. In particular, the regression algorithms are compared in this paper to predict potential problematic conditions of die-casting machines. It is natural that the number of defective condition data is much smaller than that of normal in the real-life dataset. It is caused that most defect rates of production lots are zero, and few defect rates exceed zero. This may make it difficult to predict for the defect rate for product lots.

The remainder of the paper is organized as follows. Section 2 introduces the quality prediction in manufacturing process. In Section 3, the comparative experiments of the regression algorithms are then detailed. Finally, we conclude the paper in the last section.

2. **Quality Prediction in Manufacturing Process.** In this paper, we deal with a prediction problem of defect rates in manufacturing processes, specifically the building of manufacturing quality regression models from historical process condition data. To this end, we collect the data related to manufacturing process conditions which can be gathered from various sensors and the defect rates which can be summarized from the result of product quality inspection.

2.1. **Problem and purpose.** We assume that the manufacturing process conditions cannot be mapped exactly to each product item, but it can be mapped to a production lot according to expert opinion. In other words, in this paper, the quality condition prediction problem is designed in the level of production lot, not a single. Under the assumption, we consider three kinds of data based on production lot. First, process condition data includes information of manufacturing process in time-series such as die-casting pressure, speed and injection speed. Second, lot information data contains basic information of production lots, such as the number of production items, and timestamps of lot production. Third, lot defects data contains the number of defects. These three data are joined into a single dataset in order to perform quality analysis using regression algorithms. The main purpose of this research is to compare the regression algorithms which predict defect rates of lot-level when a certain condition of manufacturing process is given.

2.2. **Data preprocessing.** We conducted two kinds of data preprocessing for the comparative experiment. The first preprocessing is the time-series data representation of manufacturing process conditions by using the statistics representatives such as maximum, minimum, mean, and deviation values of time-series data [8]. The second preprocessing is the notating of the lot-level defect rates according to the lot size and the number of defects to each defect type. For example, in cases that 180 items were produced and 36 items were defected in the certain lot, the lot-level defect rate 0.2 is notated.

3. **Experiments.** To compare the performance of regression algorithms for predicting defect rates among production lots according to the defect type, a real-life dataset from die-casting manufacturing process was used in this research. The dataset will be introduced with simple exploration results for the dataset, and the performance of three types of regression algorithms, tree-based, non-linear, and linear regression algorithms, will be presented for the given dataset.

3.1. **Dataset.** In this paper, we collected three kinds of data such as process condition data, lot information data, and lot defects data from a die-casting company in South Korea. The data recorded for 3 months from September 16, 2015 to December 10, 2015. Three kinds of data were joined to the one dataset for the experiments. The joined data contains lot information and defect rate of each defect type, and statistical values of several condition data in the die-casting process. In the dataset, there are various types of defect such as porosity, trimming, and crack. Moreover, as mentioned in previous section, the lot-level defect rates are included to the dataset by preprocessing and the lot-level defect rates become target variables of the experiments.

3.2. **Exploration.** Exploring data is an important first step in data analysis to help understand the kind of information their dataset contains [9]. We first tried to explore the dataset in order to take an insight for the comparative experiments. Figure 1 shows the example of scatter plots of the process condition such as Maximum_Speed, Minimum_Speed, Mean_Speed, Maximum_InjectionPressure, Minimum_InjectionPressure, and Mean_InjectionPressure versus the porosity defect rate. From the exploration, we found that the defect rates which are higher than zero are distributed on specific range of each

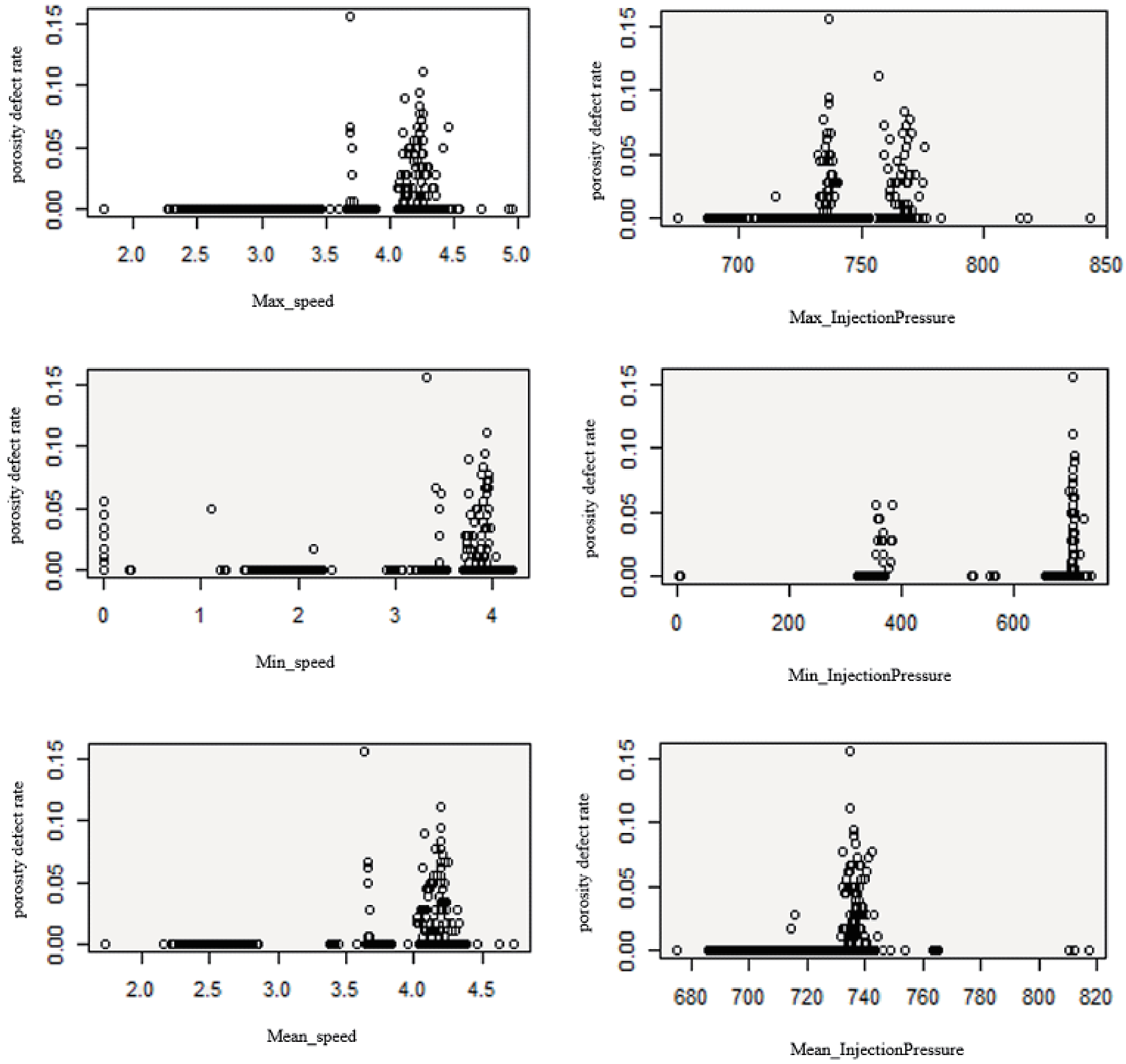



Figure 1. The example of scatter plots of the several process condition versus the porosity defect rate

process condition and thereby we could expect that the nonlinear and the tree-based regression algorithms may outperform the linear regression algorithms.

3.3. **Results.** In the next step of the exploration, we conducted the experiments of comparison of regression algorithms using the die-casting dataset. The regression models can be divided into three categories, i.e., linear regression, nonlinear regression and tree-based regression [10]. The linear regression algorithms mainly use the linear functions in the whole or the parts of the prediction models, while nonlinear regression models consider nonlinear functions as well. In addition, the tree-based regression algorithms utilize decision tree ensembles, which contain multiple decision tree models as base classifiers to complement the accuracy with one another.

The names of each tree-based models, the nonlinear models and linear models are depicted with "_(T)", "_(N)" and "_(L)" each at the end of the letters. In the categories, the performance of 22 existing regression algorithms was compared such as eXtreme Gradient Boosting (xgbTree_(T)), Tree-Based Ensembles (treeEnsemble_(T)), Partial Least Squares (pls_(L)), Linear Regression (lr_(L)), the lasso (lasso_(L)), k-Nearest Neighbors (knn_(N)), k-Nearest Neighbors (kknn_(N)), Bagged MARS (bgmars_(N)), and Neural

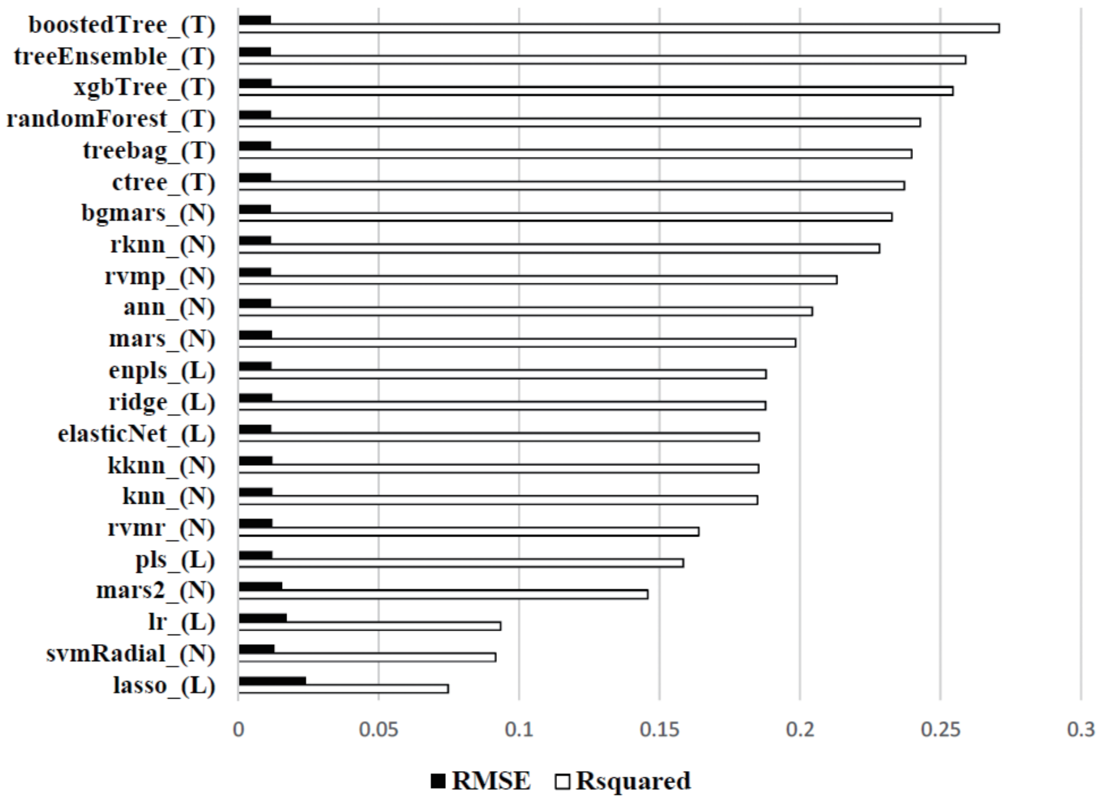
Network (ann_(N)). The experimental results summarize after five times of 5-folded cross-validation using caret package in R. We performed two type of experiments. In one experiment, feature selection was not used and all 92 variables were used to construct models. On the contrary, in the other experiment we conducted feature selection and finally 46 variables were used to construct prediction models.

We can compare the performance of 22 regression models through two evaluation measures, i.e., R-square and root mean square error (RMSE) as presented in Table 1 and Figure 2. We can interpret the comparison experiments as three results. First, the group of high-performance models are tree-based models such as Boosted Tree, Tree-Based Ensembles, eXtreme Gradient Boosting, Random Forest, Bagged CART and Conditional Inference Tree. Second, the experiment with feature selection has almost the same performance with the experiments without feature selection, in that R-squared and RMSE

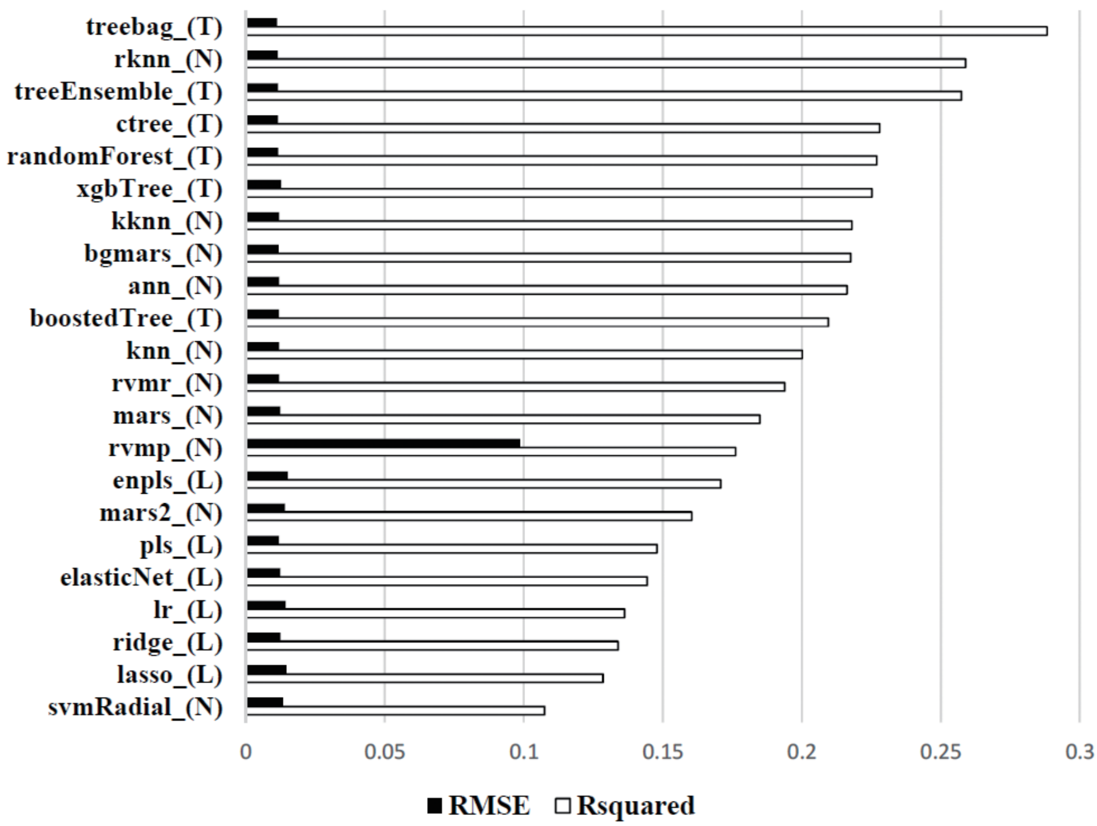
Table 1. Performance comparison among regression models for porosity defect rates

| | | w/o Feature Selection | | with Feature Selection | |
|---|---|---|---|---|---|
| Categories | Model | R-square | RMSE | R-square | RMSE |
| Tree-based models | *treebag_(T)* | 0.2397086 | 0.0117292 | 0.2882097 | 0.0112526 |
| | *treeEnsemble_(T)* | 0.2590294 | 0.0116250 | 0.2574077 | 0.0115856 |
| | *ctree_(T)* | 0.2371649 | 0.0117243 | 0.2280596 | 0.0115889 |
| | *randomForest_(T)* | 0.2428953 | 0.0117979 | 0.2269495 | 0.0117040 |
| | *boostedTree_(T)* | 0.2708933 | 0.0116436 | 0.2094885 | 0.0119232 |
| | *xgbTree_(T)* | 0.2544310 | 0.0120089 | 0.2252582 | 0.0127002 |
| Nonlinear models | *rknn_(N)* | 0.2283644 | 0.0117419 | 0.2589163 | 0.0114580 |
| | *bgmars_(N)* | 0.2328034 | 0.0116850 | 0.2176155 | 0.0118139 |
| | *ann_(N)* | 0.2044004 | 0.0117632 | 0.2162825 | 0.0119725 |
| | *kknn_(N)* | 0.1852165 | 0.0122889 | 0.2180904 | 0.0119777 |
| | *rvmr_(N)* | 0.1640064 | 0.0123030 | 0.1938306 | 0.0119803 |
| | *knn_(N)* | 0.1847791 | 0.0123097 | 0.2001377 | 0.0119807 |
| | *mars_(N)* | 0.1985021 | 0.0120273 | 0.1849109 | 0.0123934 |
| | *svmRadial_(N)* | 0.0916600 | 0.0129075 | 0.1074914 | 0.0133105 |
| | *mars2_(N)* | 0.1457509 | 0.0157347 | 0.1604124 | 0.0141249 |
| | *rvmp_(N)* | 0.2131077 | 0.0117306 | 0.1762433 | 0.0986903 |
| Linear models | *pls_(L)* | 0.1584858 | 0.0122445 | 0.1479065 | 0.0119137 |
| | *elasticNet_(L)* | 0.1854324 | 0.0118584 | 0.1444601 | 0.0123880 |
| | *ridge_(L)* | 0.1877266 | 0.0121264 | 0.1338955 | 0.0125181 |
| | *enpls_(L)* | 0.1879076 | 0.0120135 | 0.1707952 | 0.0151559 |
| | *lr_(L)* | 0.0934458 | 0.0173461 | 0.1362509 | 0.0142528 |
| | *lasso_(L)* | 0.0747073 | 0.0240760 | 0.1285417 | 0.0145402 |

*Note.* eXtreme Gradient Boosting (*xgbTree_(T)*), Tree-Based Ensembles (*treeEnsemble_(T)*), Bagged CART (*treebag_(T)*), Random Forest (*randomForest_(T)*), Conditional Inference Tree (*ctree_(T)*), Boosted Tree (*boostedTree_(T)*), Support Vector Machines with Radial Basis Function Kernel (*svmRadial_(N)*), Relevance Vector Machines with Radial Basis Function Kernel (*rvmr_(N)*), Relevance Vector Machines with Polynomial Kernel (*rvmp_(N)*), Random k-Nearest Neighbors (*rknn_(N)*), Multivariate Adaptive Regression Splines (*mars2_(N)*), Multivariate Adaptive Regression Spline (*mars_(N)*), Bagged MARS (*bgmars_(N)*), Neural Network (*ann_(N)*), k-Nearest Neighbors (*knn_(N)*), Weighted k-Nearest Neighbors (*kknn_(N)*), Ensemble Partial Least Squares Regression with Feature Selection (*enpls_(L)*), Elasticnet (*elasticNet_(L)*), The lasso (*lasso_(L)*), Ridge Regression (*ridge_(L)*), Partial Least Squares (*pls_(L)*), Linear Regression (*lr_(L)*).




(a) Experiments without feature selection




(b) Experiments with feature selection

FIGURE 2. Performance comparison among regression models for porosity defects

values are not improved. Hence, regression models using 46 variables are superior models because it has the same performance and is more simple. Third, after feature selection, performance of only nonlinear models is improved and the performance of the other models is barely improved.

4. **Conclusions.** In this paper, we dealt with the quality conditions prediction problem in the lot-level based on manufacturing condition data. Furthermore, we carried out experiments to compare the performance among regression algorithms. Representative regression algorithms such as linear regression, non-linear regression and tree-based regression are compared to predict the defective rates for production lots. A real-life dataset of die-casting manufacturing process is used to compare the performance of the regression algorithms. The experimental results show that tree-based models generally outperform linear and non-linear regression algorithms. In spite of the experiment with feature selection, the results present similar performance with the experiments which do not include the feature selection as well. It is expected that the best models could be adopted to implement the quality prediction systems for smart factory. If such valuable implementation is developed, the system can be utilized to improve the product efficiency and quality in real-world.

**Acknowledgement.** This work was supported by the Smart Factory Advanced Technology Development Program of MOTIE/KEIT (No. 10054508).

## REFERENCES

[1] K. L. Lueth, Will the industrial Internet disrupt the smart factory of the future? *IOT Analytics*, http://iot-analytics.com/industrial-internet-disrupt-smart-factory/, 2016.
[2] G. Köksal, I. Batmaz and M. C. Testik, A review of data mining applications for quality improvement in manufacturing industry, *Expert Systems with Applications*, vol.38, no.10, pp.13448-13467, 2011.
[3] D. Zuehlke, SmartFactory – Towards a factory-of-things, *Annual Reviews in Control*, vol.34, no.1, pp.129-138, 2010.
[4] W.-C. Chena, P.-H. Tai, M.-W. Wang, W.-J. Deng and C.-T. Chen, A neural network-based approach for dynamic quality prediction in a plastic injection molding process, *Expert Systems with Applications*, vol.35, no.3, pp.843-849, 2008.
[5] H. W. Cho, Enhanced real-time quality prediction model based on feature selected nonlinear calibration techniques, *International Journal of Advanced Manufacturing Technology*, vol.78, no.1, pp.633-640, 2015.
[6] A. Sata and B. Ravi, Bayesian inference-based investment-casting defect analysis system for industrial application, *International Journal of Advanced Manufacturing Technology*, 2016.
[7] G. Zhang, J. Li, Y. Chen, Y. Huang, X. Shao and M. Li, Prediction of surface roughness in end face milling based on Gaussian process regression and cause analysis considering tool vibration, *International Journal of Advanced Manufacturing Technology*, vol.75, nos.9-12, pp.1357-1370, 2014.
[8] A. A. J. Bagnall, C. Ratanamahatana, E. Keogh, S. Lonardi and G. Janacek, A bit level representation for time-series data mining with shape based similarity, *Data Mining and Knowledge Discovery*, vol.13, no.1, pp.11-40, 2006.
[9] M. C. F. De Oliveira and H. Levkowitz, From visual data exploration to visual data mining: A survey, *IEEE Trans. Visualization and Computer Graphics*, vol.9, no.3, pp.378-394, 2003.
[10] M. Kuhn and K. Johnson, *Applied Predictive Modeling*, Springer, New York, 2013.