World Academy of Science, Engineering and Technology
International Journal of Industrial and Manufacturing Engineering
Vol:2, No:12, 2008

Defect Cause Modeling with Decision Tree and Regression Analysis

B. Bakır, İ. Batmaz, F. A. Güntürkün, İ. A. İpekçi, G. Köksal, and N. E. Özdemirel

***Abstract*—**The main aim of this study is to identify the most influential variables that cause defects on the items produced by a casting company located in Turkey. To this end, one of the items produced by the company with high defective percentage rates is selected. Two approaches-the regression analysis and decision trees-are used to model the relationship between process parameters and defect types. Although logistic regression models failed, decision tree model gives meaningful results. Based on these results, it can be claimed that the decision tree approach is a promising technique for determining the most important process variables.

***Keywords*—**Casting industry, decision tree algorithm C5.0, logistic regression, quality improvement.

I. Introduction

ONE of the important issues in manufacturing processes is to determine the most influential parameters that cause defects on the items produced. Relationships between process parameters and such binary or nominal outcomes are usually modeled by using one of the traditional techniques, logistic regression approach. However, manufacturing processes are usually so complex that traditional statistical techniques or data management tools are not sufficient to extract this information. In order to manage this problem, data mining approaches [1], [2] found to be useful in other complicated areas such as customer relationship management (CRM) can be used. For the decision-makers, the ease of interpretation of results derived from analysis is as important as the predictive power of the models developed. Thus, decision trees are one of the most commonly used data mining techniques to practically solve classification and prediction problems. They have tree shaped structures in which construction of trees is simple and unlike the logistic regression models, decision tree results can be easily understood by the users. The most common types of decision tree algorithms used in the literature [3]-[6] are CART, C4.5 (or 5.0), and CHAID. In this study, we have used logistic regression and C5.0 to model relationships between process parameters and defect types for a specific product produced by the casting company.

Manuscript received October 15, 2006. This work was supported by TÜBİTAK 105M138 and Scientific Research Project 2006-07-02-06 of Middle East Technical University.

B. Bakır is with the Information Systems Department of Informatics Institute, Middle East Technical University, Ankara, 06531, Turkey (phone: +90 312 210 3739; fax: +90 312 210 3745; e-mail: bbakir@metu.edu.tr).

İ. Batmaz is with the Statistics Department, Middle East Technical University, Ankara, 06531, Turkey (e-mail: ibatmaz@metu.edu.tr).

F. A. Güntürkün is with the National Productivity Center, Ankara, Turkey (e118397@metu.edu.tr).

A. İ. İpekçi is with the Scientific Decision Support Center, Turkish General Staff, Ankara, 06550, Turkey (aipekci@tsk.mil.tr).

G. Koksal and N. E. Ozdemirel are with the Industrial Engineering Department, Middle East Technical University, Ankara, 06531, Turkey (e-mail: koksal@ie.metu.edu.tr, e-mail: nurevin@ie.metu.edu.tr).

II. Casting Defect Problem and Data Description

The flow of whole casting process is shown in Fig. 1. Defect of various types occur typically due to parameter settings in the melting and casting stages.

One of the quality objectives of the company is to reduce the percentage of defective items by identifying and optimizing the most important process parameters. This is typically achieved by analyzing data collected by designed experiments. However, before such experimentation, it is necessary to determine the most significant factors involved in the process.

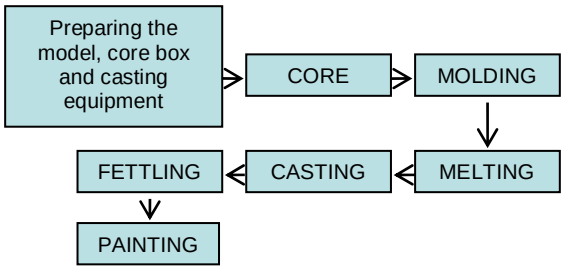

Fig. 1 Manufacturing Line

The company has provided us with the data for a particular product (a cylinder head), which has high percentage of defectives collected during the first five-month production period of 2006. Data used in this study are observational data and particularly collected from three subsequent processes, which are core, molding and melting. The company records values of certain parameters hourly, daily or weekly to monitor the production processes without conducting any specific data analysis. During the production period of a batch, values of input parameters are recorded by sampling; therefore, these values are taken as the same for all individual items produced in that batch. For that reason, every individual item is associated with the average values of the batch the item belongs to. Since some

values of the parameters have not been recorded by the company, these values are treated as missing in the analysis of data. Although, ten different defect types were recorded, only two most important and frequently observed defect types have been considered in this study. Thus we have a nominal response variable having three levels. Two of them, coded as 1 and 2 represent the two defect types selected and the one coded as 0 means neither of these defects has been observed. There are 36 continuous type input variables measured on 809 defective items produced.

In this study, we have analyzed these records by using both traditional (logistic regression) and data mining (decision tree) approaches and discussed if the results can help the company reduce defects on the items.

III. Regression Analysis and Results

One of the traditional techniques used to determine the most influential variables involved in a production process is regression analysis. Therefore, in this study, we have first tried to develop a regression model which relates defect types to input factors. Since the response variable is of nominal type with three categories, multinomial logistic regressing approach [7], [8] is applied using Clementine 10.1 which is the data mining software of SPSS.

We have used the stepwise regression technique to develop a multinomial model involving all main effects and two-way interactions of 36 input variables. Although the fitted model is statistically significant (p-values for Pearson and Deviance are 1.0; p-value for G is 0.0), none of the parameter estimates are found to be significant.

Since different defect types may be caused by different factors involved in the production process, we have also modeled defect types separately by using binary logistic regression. In the analysis of these models, we face the problem of complete separation which leads to either infinite or non-unique maximum likelihood parameter estimates. One of the possible reasons of this problem is sensitivity of classification to the relative sizes of the two response groups which is also the case for our data set [7]. As a result, it may be concluded that the data does not fit the model [9]. Alternatively, as suggested by some studies including [10], [11], arrangements can be made on data sets to overcome this problem. Since this approach requires more elaborate solutions, they have been left out of the scope of this study.

IV. Decision Tree Approach and Results

A decision tree is a simple tree-shaped structure where each internal node represents a test on one attribute, arcs show the results of a test and leaf nodes reflect classes. They are easy to understand and can be easily converted to a set of rules. Moreover, they can classify both categorical and numerical data and require no priori assumptions about the data. Because of the advantages listed above, the decision tree approach is extensively utilized for both classification and prediction purposes.

In this study, we have used C5.0 algorithm [12], which is an improved version of C4.5. The algorithm C5.0 is a commercial product designed by RuleQuest Research Ltd Pty to analyze huge databases. We have used Clementine 10.1 to implement this algorithm.

The nominal response variable is used to develop the decision tree model. The data set is randomly divided into training and testing sets with the approximate proportions of 70% and 30%, respectively. During the training session, boosting method with 10 trials is used to improve the accuracy of the model. After constructing the tree, global pruning is performed with 75% pruning severity to avoid overfitting. Minimum 5 records are allowed in leaf nodes. Estimated accuracy for the final model is found to be 92.15% for the training set. According to the results presented in Table I, the model suggested correctly classifies 91.93% of the testing data. As a result, it can be said that the performance of the model on the test data set is as good as the performance on the training data set.

The decision tree model finds nine process variables to be influential on the response, defect types, and it also extracts ten rules associated with these significant input variables (see: Fig. 3 in appendix A). One of the extracted rules gives the important variables and threshold values for the defect type that cannot be measured by the company and that can only be determined after usage by the customers. Percentage of products returned by the customers for this reason has an increase. Therefore, the rule extracted from the decision tree model developed is valuable for the company. Confidence level of the rule is 82.8% and it is valid for 164 records. Evaluation graph in Fig. 2 shows that performance of the model for this problematic defect type is very close to the best model indicating perfect confidence.

TABLE I
Coincidence Matrix for Predicted Categories

Training	**0**	**1**	**2**
0	33	0	5
1	0	162	16
2	0	25	345
Testing			
0	15	0	1
1	0	50	5
2	0	12	140

Rows show actual values.

A. Rules Extracted from the Tree

Ten rules the decision tree provides are listed below. Values in parentheses show the number of records (instances) to which the rule applies, and the proportion of those records for which the rule is true (confidence). Although confidence levels of rules 3, 7 and 9 are high, these rules may be omitted because of only few observations distinguished by the rules.

Rule 1: (17; 1.0)
IF X22 <= 14.35 AND X8 <= 35 THEN Y = 0
Rule 2: (16; 1.0)
IF X22 > 14.35 AND X27 <=4.2 AND
X35 > 0.088 THEN Y=0
Rule 3: (5; 1.0)
IF X22 <= 14.35 AND X8 > 35 AND
X30 <= 1.88 AND X2 > 23 THEN Y=0
Rule 4: (198; 0.828)
IF X22 > 14.35 AND X27 >4.2 AND
X9 <= 3.216 AND X12 > 305 AND
X19 >15.95 THEN Y=1
Rule 5: (13; 1.0)
IF X22 <= 14.35 AND X8 > 35 AND
X30 <= 1.88 AND X2 <= 23 THEN Y=2
Rule 6: (268; 1.0)
IF X22 <= 14.35 AND X8 > 35 AND
X30 > 1.88 AND THEN Y=2
Rule 7: (8; 0.875)
IF X22 > 14.35 AND X27 <= 4.2 AND
X35 <= 0.088 THEN Y=2
Rule 8: (34; 0.765)
IF X22 > 14.35 AND X27 > 4.2 AND
X9 > 3.216 THEN Y=2
Rule 9: (9; 0.889)
IF X22 > 14.35 AND X27 > 4.2 AND
X9 <= 3.216 AND X12 <= 305
THEN Y=2
Rule 10: (18; 0.778)
IF X22 > 14.35 AND X27 > 4.2 AND
X9 <= 3.216 AND X12 > 305 AND
X19 <= 15.95 THEN Y=2

V. Conclusion

The logistic regression technique has not been useful in estimating unique parameter values to identify important process variables that cause defective items. This conclusion is compatible with some of the findings in literature [13]. However, the decision tree approach has provided us with satisfactory results. This is likely to be caused by the partitioning facilitated by the tree construction. These results were presented to the quality team of the company. Some of the parameters and their respective thresholds in the model were judged meaningful, whereas some others were found to be unexpected (interesting). In addition, threshold values of the parameters provided by the model were considered to be useful in optimization of the casting process. The company decided to conduct controlled experiments on the significant model parameters utilizing the threshold values in choosing the levels of the parameters. In this sense, the decision tree analysis can be considered as a way of planning for statistical design of experiments for optimization purposes.

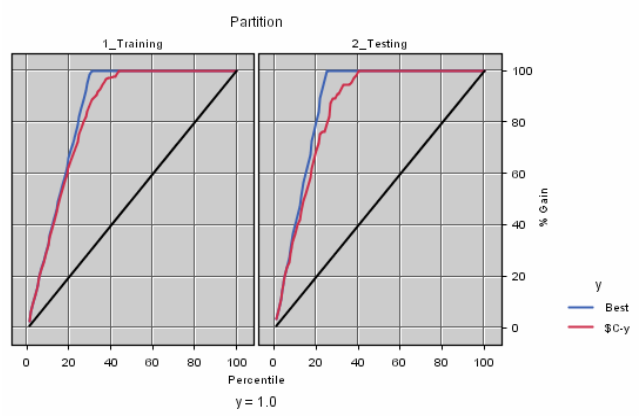

Fig. 2 Gains Chart (hit: y = 1)

Acknowledgment

We would like to thank to all team members of these projects and casting company for their contributions to this study.

References

[1] J. Han, M. Kamber, *Data Mining: Concepts and Techniques,* Morgan Kaufmann Publishers, 2001.
[2] M. H. Dunham, *Data Mining: Introductory and Advanced Topics.* Prentice Hall, 2003.
[3] B. S. Kang, S. C. Park, "Integrated machine learning approaches for complementing statistical process control procedures", *Decision Support System,* vol. 29, pp. 59-72, 2000.
[4] M. Li, S. Feng, I. K. Sethi, J. Luciow, K. Wagner, "Mining Production Data with Neural Network & CART" in *Conf. Rec. 2003 IEEE Int. Conf. Data Mining.*
[5] J. Lian, X. M. Lai, Z. Q. Lin, F. S. Yao, "Application of data mining and process knowledge discovery in sheet metal assembly dimensional variation diagnosis", Journal of Materials Processing Technology, vol. 129, pp. 315-320, 2002.
[6] D. Braha, A. Shmilovici, "Data Mining for Improving a Cleaning Process in the Semiconductor Industry", *IEEE Trans. Semiconductor Manufacturing,* vol. 15, no. 1 pp. 91–101, Feb. 2002.
[7] D. W. Hosmer, S. Lemeshow, *Applied Logistic Regression.* Wiley-Interscience Publication, 2000.
[8] D. C. Montgomery, E. A. Peck, *Introduction to Linear Regression Analysis.* Wiley, 1982, pp. 444-453
[9] P. McCullagh, "Regression models for ordinal data (with discussion)", *Journal of the Royal Statistical Society. Series B, vol. 42, pp. 109-127, 1980.*
[10] A. Albert, J. A. Anderson, "On the existence of maximum likelihood estimates in logistic models", Biometrika, vol. 71, pp. 1-10, 1984.
[11] M. C. Bryson, M. E. Johnson, "The incidence of monotone likelihood in the Cox model", Techometrics, vol.23, pp. 381-384, 1981.
[12] Data Mining Tools C5.0
http://www.rulequest.com/see5-info.html
[13] K. R. Skinner, D. C. Montgomery, G. C. Runger, J. W. Fowler, D. R. McCarville, T. R. Rhoads, "Multivariate Statistical Methods for Modeling and Analysis of Wafer Probe Test Data", *IEEE Trans. Semiconductor Manufacturing,* vol. 15, no. 4 pp. 523–530, Nov. 2002.

APPENDIX A

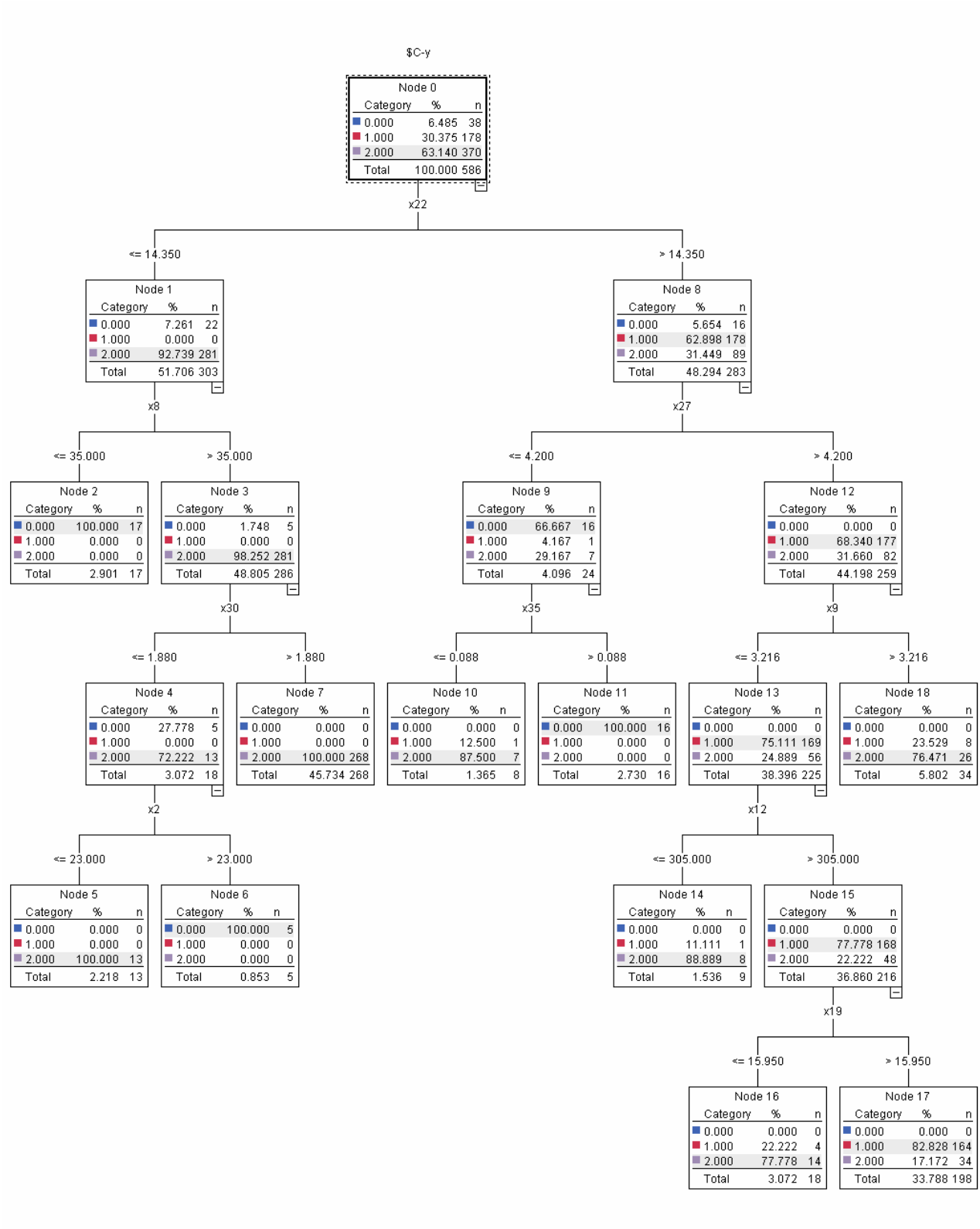

Fig. 3 Graph representation of the developed C5.0 model