STATISTICAL MODEL TO PREDICT THE PERFORMANCE OF MANUFACTURING MODULES IN SRI LANKAN APPAREL INDUSTRY

[1]NAWODH W.A.S., [2]LOKUPITIYA R., [3]SAMAYAWARDENA D.

[1,2]Department of Statistics, Faculty of Applied Sciences, University of Sri Jayewardenepura, Gangodawila, Nugegoda, Sri Lanka
[3]ABC (Pvt) Ltd, Sri Lanka
E-mail: [1]wasnawodh@gmail.com, [2]lokupitiya@sjp.ac.lk

Abstract - With the development of technology and increasing competition between manufacturing firms in the global market, it has become essential for organizations to transform the manufacturing process to predictive manufacturing. Especially in Sri Lankan apparel industry, with increasing complexity in manufacturing modules as well as to keep the competitiveness against other global players, going towards predictive manufacturing is crucial. One of the key elements of predictive manufacturing is to identify the lower performing manufacturing modules in the factory so that management can take proactive actions to improve the performance of this volatile environment. As a solution, this study was conducted to develop a statistical predictive model to predict whether a manufacturing module will achieve its given target at the end of the day. Since the depended variable is binary, LogisticRegression model, Naïve Bayesianand Support Vector Machine algorithms were chosen. Based on domain knowledge expertise,predictor variables for the study were identified. To develop the model, data form five months were collected from the chosen factory and analysis was carried out using the R programming language. Collected data of the factory consisted with class imbalance problem and due to this, Mathews Correlation Coefficient was used for model evaluation. Through model improvement procedures, we were able to improve the model performance and final overall accuracy of the fitted model was 78%.

Keywords - Performance prediction, Logistic Regression, Manufacturing Prediction, Data Mining

I. PURPOSE AND BACKGROUND

In the present global economy, embracing emerging technologies like advance analytics and internet of things (IOT) to manufacturing industry has become a necessity for the organizations and industries. With the emergence of these, to retain the competitiveness in global economy, organizations are pushing themselves towards transforming their manufacturing to predictive manufacturing (Lee, et al., 2013).This phenomenon is also applicable to Sri Lankan apparel industry. As Sri Lankan apparel industry is a key contributor to the country's export market, it is crucial to keep its competitive edge over others by transforming shop-floor environment to a predictive manufacturing environment. Another reason isdue to drastic changes in consumer behavior complexity of the manufacturing modules in the industry has increased and continues to be so. Due to this phenomena, number of garment styles coming to the factory are increasing while decrease in customer lead time; time from customer placing the order to the time customer receiving the produced order. As a result of this situation, unlike in the past, time which can be allocated for the learning for the team members and adaptation time for the manufacturing plants is decreasing over the time. To overcome these industrial challenges, it is essential to transform manufacturing to a predictive manufacturing.

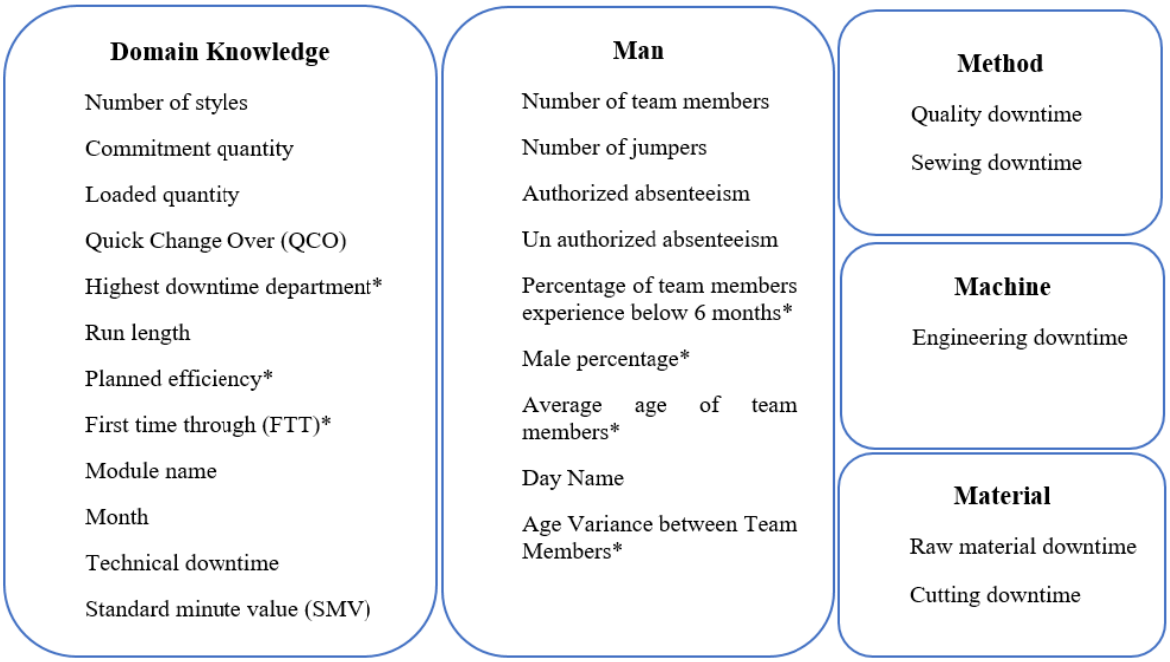
Using manufacturing execution systems (MES), enterprise resource planning (ERP) systems and other management systems linked with shop-floor, real time data capturing is now possible with manufacturing organization. Keeping these data in a database adds no value to the organization and it is possible to convert these data to a valuable source with the support of good analytical tools, which has been argued in many literature (Stein and Flath, 2017). With proper application of analytical tools, predictive model for shop-floor can be developed so that management can take data driven strategic and operational decisions. In relation to the apparel industry, identifying lower preforming manufacturing modules is an integral part of predictive manufacturing. This identification is crucial for the management, so that they can take decisions proactively to improve their performance.

Objective of this study is to develop a statistical predictive model to predict the performance of manufacturing modules in apparel industry using real world data rather than simulated data. Novelty has been brought to this study by application of statistical models to predict the performance of manufacturing modules in Sri Lankan apparel industry using real world data. The performanceis measured as a binary response variable which is derived from a key performance indicator (KPI) known as plan to performance (PTP). It is defined as whether a manufacturing module achieved its target or not. Equation (1) elaborates how PTP KPI is calculated.

$$\text{PTP KPI} = \frac{\text{Number of units producted by the module in the day (Performed)}}{\text{Number of units planned to produce by the module in the day (?lanned)}} \times 100 \quad (1)$$

II. METHODOLOGY

Initially variables for the study were identified using the 4M approach (Man, Method, Machine, and Material) and the domain knowledge of industry expertise. In this process, availability of data and data completeness conditions were considered before selecting a variable. The identified variables can be separate into two groups as primary variables and secondary variables. Primary variables are the variables which have been directly taken from the shop-floor data while secondary variables are the derived variables from shop-floor data. Table 1 displays the identified variables in 4M approach.



* Secondary Variables

Table 1. Independent variables of the study

Once the variable identification is completed, data is collected from fourteen different data sources for a period of five months staring from June to October. The chosen manufacturing plant for this study consists of twenty manufacturing modules which consist of 400 employees. For a day, twenty data points were collected, such that each row explains behavior of a given manufacturing module on a given day. Collected data for above variables consist of twenty continuous variables and six categorical variables and the depended variable is a binary variable which is labeled as 1, if PTP is greater than or equal to 100%, else labeled as 0. In collected data, only 25% data points have achieved its target indicating that dependent variable is not balanced. According to definition by Li and Sun (2012), since minority class has a proportion constitute less than 35% from the total sample, this sample is an imbalanced one. Because of that, this study consists of a class imbalance problem. When class imbalance problem is present in the data, data mining algorithms tend to perform higher predictive accuracy on majority class while poor predictive accuracy on minority class (Thammasiri, et al., 2013). Due to this it is important to find an appropriate model evaluation matrix.

For model evaluation in classification problems, confusion matrix is used. Confusion matrix is an table which represents number of correct prediction and incorrect predictions and for a binary classification problem it is a two-way table as given in Table 2.

		Predicted Results	
		Yes	No
Actual Results	Yes	True Positive (TP)	False Positive (FP)
	No	False Negative (FN)	True Negative (TN)

Table 2. Confusion Matrix for Binary Classification

True Positive (TP) represents the number of data points predicted correctly when actual result is “Yes” or “1”; True Negative (TN) represents the number of data points predicted correctly when actual results is “No” or “0”. False Negative (FN) represents the number of data points predicted incorrectly when actual result is “No” or “0”; False Positive (FP) represents the number of data points predicted incorrectly when actual results is “Yes” or “1”. From the above confusion matrix following measures are used to evaluate and compare the models.

Even though Accuracy, Recall, Precision and F-Score are commonly used in model evaluation, when it comes to an imbalance classification problem, Boughorbel et al. (2017) argues that using Matthews correlation coefficient (MCC) is more suitable for imbalance classification. Similar argument was made by Stein and Flath (2017) following (Powers, 2011). Equation (2) – (7) are the equations to calculate Accuracy, Recall, Precision, F-Score and MCC, respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$F - Score = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{5}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{7}$$

Multicollinearity was checked between independent variables in this study as Logistic Regression (LR) model assumes there is no multicollinearity between independent variables. To identify multicollinearity, generalized variance inflations Factors (GVIF) is calculated and any variable which has GVIF greater than five is considered to have multicollinearity.

Since the dependent variable is binary initially, model development is carried out using LR model. Initially LR model was fitted to the original data and then to the cleaned data. As independent continuous variables are in different scales and different units, each variable was standardized. Due to having many independent variables, to improve the model, bi-directional step wise feature selection technique and Principle Component Analysis (PCA) technique were considered as feature reduction techniques to improve the model. Naïve Bayesian (NB) and Support Vector Machine (SVM) classifiers were also implemented after this process.

III. RESULTS AND DISCUSSION

Initially, for the collected data LR was applied naively without any data clean or variable transformations; records with missing values were removed from the data sheet. By applying so, model (LROD)obtained a MCC of 0.22. As there is room for improvement, data cleaning and missing value imputation processes were carried out. For missing value imputation, multivariate imputation by chained equations (MICE) were utilized. Using the missingness plot, it was identified that 3% of data is missing and removing those would end up removing considerable number of records. Using the domain knowledge, some extended missing values were filled out and remove unnecessary data rows. Even then, a considerable amount of missing data was present specially with respect to absenteeism. MICE was applied to impute the missing values of absenteeism; total missing data was reduced to 1%.

Once the cleaning process was completed, LR model was applied to clean data (LRCD) and obtain a MCC of 0.30. To improve the model performance further, a new continuous variable was derived from six different downtimes to replace them which represents the total module downtime of each manufacturing module. Applying LR regression to the clean data which includes new derived variables (LRCDV1), obtained a MCC of 0.32. A new categorical variable was derived from the variable date name, to replace it which represent the nature of the day. The LR regression which was applied to the cleaned data with both derived variables (LRCDV2) was able to obtain an MCC of 0.28. Even though adding the second new variable led to reduce the model performance, it was kept and continued the model improvement process. By observing the cleaned data and using domain knowledge, four variables which consist of outliers were identified; these outlier records were removed. With the removal of outliers, LR regression model (LRCDV2OR) performance increased up to 0.34.

Next, feature selection processes were carried out to identify most prominent features for the predictive model. For this, both stepwise method and principle component analysis (PCA) were applied. Out of these two techniques initially, PCA was applied and obtained ten principle components (PCs) which explains 80% of total variance of the original variables. Without directly applying LR model to PCs, model was applied to both PCs and Categorical data (LRCDV2ORPCA) and obtain a MCC of 0.40. To further improve this model stepwise procedure was applied to identify most prominent features out of PCs and categorical data (LRCDV2ORPCASW); obtained an MCC of 0.41. Since no further improvement can be taken from PCA approach, step wise approach was tested. By applying forward and backward stepwise procedure to the LR model (LRCDV2ORSW), obtained a MCC of 0.38. By looking at the selected features, new features were added to the above model and few features were exempts from the model using domain knowledge (LRCDV2ORDK) and obtain a MCC of 0.43 from the new model. To further improve the LR model, eight continuous variables was converted to categorical variables by grouping. This was carried out by looking at the distributions and use of domain knowledge (LRCDV2ORDKCT). With this conversion, obtained a MCC of 0.48. by only taking

the significant variables to the LR model (LRCDV2ORDKCTSG), MCC of the new model was 0.46. from this is evident that feature selection through stepwise approach is more suitable than PCA approach.

Figure 1 displays the improvement of the LR model through stepwise approach. For the final selected variables; NB, SVM was implemented to compare the performance between algorithms. Table 3 displays the accuracy, precisions, recall, F-Score and MCC scores of each model developed in the study.

Algorithm	Accuracy	Precision	Recall	F-Score	MCC
LROD	0.75	0.53	0.24	0.34	0.22
LRCD	0.78	0.61	0.29	0.39	0.30
LRCDV1	0.79	0.66	0.28	0.39	0.32
LRCDV2	0.78	0.62	0.24	0.34	0.28
LRCDV20R	0.76	0.74	0.29	0.41	0.34
LRCDV20RSW	0.74	0.64	0.48	0.55	0.38
LRCDV20RDK	0.76	0.69	0.51	0.59	0.43
LRCDV2ORDKCT	**0.78**	**0.71**	**0.56**	**0.63**	**0.48**
LRCDV20RDKCTSG	0.77	0.69	0.55	0.61	0.46
NB	0.73	0.58	0.68	0.63	0.42
SVM (C=0.5, Sigma = 0.01422941)	0.73	0.81	0.22	0.35	0.32

Table 3.Model evaluation scores for algorithms

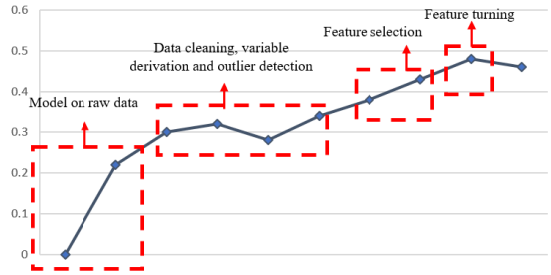


Figure 1. MCC improvement of the LR model

From Table 3, the best model with high predictability is the model LRCDV2ORDKCT. The final LR model equation is displayed in equation (8) and coefficients value of the equation (8) is in the Table 4.

$$\ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 \times \text{Combine Status} + \beta_2 \times \text{Number of Styles} + \beta_3 \times \text{Comitment Qty(Categorical)} + \beta_4 \times \text{Loading Quantity (Categorical)} + \beta_5 \times \text{Downtime (Mins; Categorical)} + \beta_6 \times \text{Absenteeism (Categorical)} + \beta_7 \times \text{Number Jumpers (Categorical)} + \beta_8 \times \text{Age Variance} + \beta_9 \times \text{Day Label} + \beta_{10} \times \text{Module} + \beta_{11} \times \text{Run Length (Categorical)} + \beta_{12} \times \text{Planned Efficiency} + \beta_{13} \times \text{FTT(Categorical)} + \beta_{14} \times \text{SMV(Categorical)} \quad (8)$$

where P = probability of a module achieving PTP 100% or above in a day.

Coefficient	Value	Coefficient	Value	Coefficient	Value
(Intercept)	-6.10***	Module LL02	1.76***	Module LL16	2.79***
Combine Status (Combine with 17)	-14.75	Module LL03	1.85***	Module LL17	3.23***
Combine Status (Combine with 18)	-0.64	Module LL04	2.32***	Module LL18	2.33***
Combine Status (Combine with 12)	0.33	Module LL05	1.22*	Module LL19	1.56**
No of Styles	-0.09	Module LL06	0.87.	Module LL20	1.39*
Commitment quantity (> 850)	-0.34.	Module LL07	2.02***	Run Length (>55 Days)	1.22***
Loading Quantity (> 915)	0.39*	Module LL08	1.65***	Run Length (23 < Run length ≤ 55)	0.55***
Down Time (> 6 Hours)	-0.45	Module LL09	1.63***	Planned Efficiency	2.63**
Down Time (3 < DT ≤ 6 Hours)	-0.49**	Module LL10	1.52**	FTT (> 0.94)	2.12***
Absenteeism (>3)	-0.30	Module LL11	1.03*	FTT (0.86 < FTT ≤0.94)	1.05**
Jumps (>3)	-0.87	Module LL12	1.95***	SMV (8.4 < SMV ≤ 10.7)	0.07
Age Variance	0.00	Module LL13	1.92***	SMV (10.7 < SMV ≤ 12)	1.15***
Day Labe (Day Before Holiday)	-0.08	Module LL14	2.01***	SMV (12 < SMV ≤ 17.2)	-0.70
Day Labe (Day after Holiday)	-0.48**	Module LL15	2.62***	SMV (SMV > 17.2)	-0.45

Significant codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

Table 4.Coefficient values of final LR model

IV. CONCLUSION

With collected data from MES, ERP and other systems which is connected to shop-floor, it can be concluded that predictive model can be developed to the shop floor of the apparel industry. In this study, it has proven that application of white box algorithms like, LR and NB; is able to get satisfactory results. It was also observed that mere applying algorithms to data set does not produce any good results, but through continuous improvement in the algorithm by missing value imputing and identifying anomalies, does improve the algorithms performance. Further the algorithm performance can be improved by applying suitable feature selection techniques. Throughout this study, it was observed that having the domain knowledge plays an important role from the point of variable identification till end of the model development process. LR algorithm is considered as a simple white box algorithm, yet in this study, it has been proven that it performs better than NB and SVM. By further studying shop-floor environment and adding more critical variables may improve the algorithms performance further and application of black box algorithms to apparel shop-floor predictive model will improve the predictability further.

REFERENCES

[1] Boughorbel, S., Jarray, F. & El-Anbari, M., 2017. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLoS ONE 12(6).

[2] Lee, J., Lapira, E., Bagheri, B. & Kao, H.-a., 2013. Recent advances and trends in predictive manufacturing systems. Manufacturing Letters 1, pp. 38-41.

[3] Li, H. & Sun, J., 2012. Forecasting business failure: The use of nearest-neighbour support vectors and correcting imbalanced samples e Evidence from the Chinese hotel industry. Tourism Management 33, pp. 622-634.

[4] Powers, D. M. W., 2011. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies.

[5] Stein, N. & Flath, C., 2017. Applying data science for shopfloor performance prediction. Guimarães, Association for Information Systems.

[6] Team, A. V. C., 2018. Practical Guide to deal with Imbalanced Classification Problems in R. [Online]

[7] Available at: https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/

[8] Thammasiri, D., Delen, D., Meesad, P. & Kasap, N., 2013. A critical assessment of imbalanced class distribution problem: The case. Expert Systems with Applications.

★★★