AN ANALYSIS FOR PREDICTION OF DELAY OCCURRENCES IN BUILDING MANAGEMENT SYSTEMS IN SRI LANKA

D.S. Rodrigo[1], K.G.H.S. Peiris[2], M.A. Reyal[3], I.U. Amarawicrama[4]

[1] *Senior Lecturer, Department of Mathematics, University of Sri Jayewardenepura, Nugegoda, Sri Lanka*
[2] *Temporary Lecturer, Department of Mathematics, University of Sri Jayewardenepura, Nugegoda, Sri Lanka*
[3] *Undergraduate, Department of Mathematics, University of Sri Jayewardenepura, Nugegoda, Sri Lanka*
[4] *Undergraduate, Department of Mathematics, University of Sri Jayewardenepura, Nugegoda, Sri Lanka*

ABSTRACT

Predicting whether a delay arises in a BMS project, assists with early resource and time allocation. This study intends to provide BMS project managers and clients at an early stage of a project, an accurate yes/no on whether a delay could occur. Through descriptive statistics, variables with significance to delay occurrence are identified. The main findings from the descriptive statistics were; out of all quantitative variables, the number of points and the estimated time of a BMS project were identified as significant by dot plot diagram. From the Chi-square test for association done for all categorical variables, client type proved to be significant with a p-value of 0.037 (<0.05) for the association between the variables, client type and delay occurrence. Considering these significant variables, linear discriminant models and binary logistic models are constructed. Both types of models will be validated by k-fold validation and the most accurate model will be chosen using a classification table. The binary logistic models considered all significant variables but, the linear discriminant model was constructed by only the quantitative predictors as per the assumption of constructing such a model. Therefore, only the number of points and the estimated time of a BMS project was considered to build the linear discriminant model. The most accurate model was identified as a linear discriminant model constructed in the 3rd fold of delay occurrence vs. points and the estimated time had the best accuracy (0.8571). This model proved more accurate than all models constructed in the study. The accuracy reveals that a linear discriminant function is more accurate to predict the occurrence of a delay in BMS projects than a binary logistic model.

Keywords: - *Discriminant Analysis, Binary Logistic Regression Analysis, Building Management System, Number of Points, Client Type.*

1. INTRODUCTION

Building Management Systems (BMS) is a complete automation program that links all building components, energy uses, controls, and schedules with a central system used to monitor the energy use pattern of the building and optimize energy efficiency. More specifically, they link the functionality of individual pieces like lifts, CCTVs, Fire Alarms, ACs so that they operate as one complete integrated system. This study was conducted using the data obtained by several companies in Sri Lanka that offer BMS installation. These companies play a crucial role in equipping several renowned buildings in Sri Lanka with BMS. This field differs from any traditional construction and civil engineering projects, thereby lacking a method to predict delay occurrences in BMS projects.

2. METHODOLOGY

The target population of this study is BMS projects implemented in Sri Lanka. As the field of BMS is growing in Sri Lanka, the population size ranges from 60 to 70 projects. Also, these projects were conducted by several BMS supplier companies, which are mutually homogeneous and internally heterogeneous groups. Hence the sufficient number of BMS projects implemented by several BMS supplier companies in Sri Lanka were selected using a two-stage cluster sampling. A practical sense of BMS was achieved by inspecting specific ongoing BMS projects and

conducting discussions with project managers and experienced employees of selected BMS suppliers. From these discussions, many quantitative and categorical variables were identified as parameters, which comprises the possibility of influencing delay occurrence.

Quantitative variables include the number of points (the sensors and actuators in a BMS project), the estimated time for a project and actual time for the completion of the project. The categorical variables were Client type (government sector /private sector), building type (HiRISE/spread), sector (factory/hotel/office), occupancy (construction field is occupied or not; yes/no), project location (within the capital city or not; yes/no) and BMS type (installation method: Chiller Management System (CMS) or BMS). Here the labor type was not identified as a parameter since the considered BMS suppliers outsource National Vocational Qualification (NVQ) qualified electricians and technicians.

Descriptive statistics is the descriptive coefficients compiling a data set that is a representation of the entire population or a sample [9] (Trochim 2004, p. 59). The collected data from BMS projects are descriptively analyzed to depict and summarize data in a meaningfully using the tool MINITAB19. Considering all variables of interest univariate analysis was performed to identify the composition of the sample. Also, a bivariate analysis was performed between the dependent variable and all explanatory variables. The Chi-square test tells the probability of independence of a distribution of data. It is used to analyze categorical data [8] (Rana and Singhal, 2015). In the bivariate analysis, Chi-square test was used to find whether there is an association between all categorical variables with the occurrence of delay. After identifying the significant parameters, the study dives into modeling.

Two different statistical techniques namely Linear Discriminant Analysis and Binary Logistic Regression Analysis were used to predict the delay occurrence. A comparison of the logistic model and the discriminant analysis will be implemented to calculate delay occurrence and the most accurate model will be chosen.

According to Alayande & Adekunle, 2015, discriminant analysis is a technique used by the researcher to analyze the research data when the dependent variable is categorical and the predictor or the independent variable is an interval in nature. The objective being is to develop functions that will discriminate between the categories of the dependent variable correctly. It enables the researcher to examine whether significant differences exist among the groups, in terms of the predictor variables [1]. The assumptions that should be satisfied to conduct a linear discriminant analysis are; the response variable should indicate the group, predictors should not be highly correlated and the equality of the covariance matrices for all groups. Bartlett's Test was used to identify the equality of the covariance matrices for all groups. Bartlett's Test is a hypothesis test used for homogeneity of variances. It determines whether a statistically significant difference exists between the variances of two or more independent sets of normally distributed continuous data [6] (National Institute of Standards and Technology International SEMATECH, 2003).

However, binary logistic regression models the relationship between a set of predictors and a binary response variable. A binary response has only two values, such as yes/no, suitable for stating whether or not a delay would occur. The binary regression model is used to understand how changes in the predictor values are associated with changes in the probability of an event occurring [3] (Hosmer & Lemeshow, 2013). In a binary logistic model, the dependent variable must be binary as same as in the discriminant analysis. The dependent variables need not be an interval, no normally distributed, no linearly related, and no equal variance within each group. This method can be used for adequate sample size and the absence of multicollinearity assist to have the best results. Multicollinearity exists when there is a correlation between the predictors in a model and the Variation Inflation Factor (VIF) estimates how much it inflated the variance of a regression coefficient [5] (Kassambara, 2018). The smallest possible value of VIF is one; in which case it is the absence of multicollinearity. As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity [4] (James et al. 2014). Hence the absence of multicollinearity could be identified by VIF.

To minimize the bias associated with the random sampling and holdout data samples, the *k*-fold cross-validation is used. In this testing method, the complete data set is randomly split into *k* mutually exclusive subsets of approximately equal size. The classification model is trained and tested *k-1* times and each time it is trained, one of the *k-1* subsets is tested and the rest *1* is taken as the testing data [7] (Olson & Delen 2008, pp. 144-145). With testing delay occurrence models, firstly the goodness of fit of the binary logistic model was tested by the Hosmer-Lemeshow (HL) test; which tells you how well your data fits the model. Specifically, the HL test calculates if the observed event rates match the expected event rates in population subgroups [2] (Glen, 2016).

Another way of evaluating the fit of a logistic regression model/discriminant function is via a classification table [10] (Zaiontz, 2018). All models were evaluated and compared by using classification tables and the most accurate model to predict the occurrence of delays was selected according to the accuracy given by classification tables.

3. RESULTS AND DISCUSSION

3.1 Descriptive Statistics

According to the descriptive analysis, approximately 60% of BMS projects were not completed during the estimated time. So, it is needed to identify the variables which could affect delay occurrences in BMS projects. According to figure 1, the chance of occurring delay was higher when there were a large number of points and the estimated time was long. When considering the delays with respect to client type, nearly all projects of the government sector has recorded delays occurrence. Occurring delays in private sector projects were approximately 55%. The occurrence of delays according to building type was 65% and 56% in HiRISE and spread respectively. Occurring delays in BMS was higher than the occurring delays in CMS. The percentage of occurring delays in BMS projects conducted in occupied buildings was greater than the percentage of occurring delays in BMS projects conducted in non-occupied buildings. Approximately 38% of BMS projects conducted for factories were delayed while 70% of BMS projects conducted for offices and hotels were delayed.

To identify the variables affecting delays in BMS projects, the Chi-square test for association was conducted between delay and all other categorical variables. According to the Chi-square test, among all the categorical variables, the most significant parameter for predicting delay occurrence is client type. The p-value of the association =0.037 < 0.05. Hence, there is enough evidence to reject the null hypothesis with a 0.05 significance level. Therefore, it can be concluded that there exists an association between delay and client type with 95% confidence. According to the dot plots, the number of points and estimated time were significant parameters for delay occurrence (figure 1).

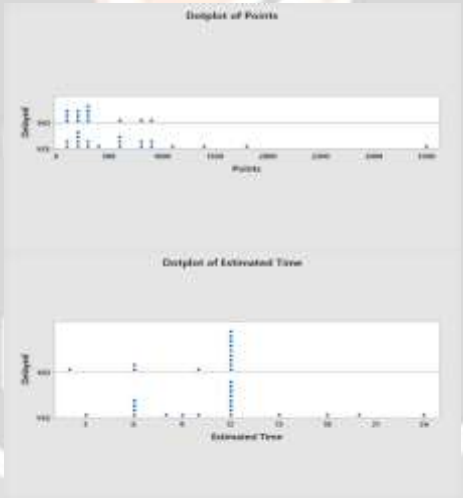

Fig -1: Dot plots of the number of points and estimated time in delayed and not delayed groups

Therefore, by considering both quantitative variables, points and estimated time, a linear discriminant analysis was conducted to predict the occurrence of delays in BMS projects and by considering categorical variable, client type as well as quantitative variables, binary logistic regression analysis was conducted to predict the occurrence of delays in BMS projects.

3.2 Discriminant Analysis

The variables, number of points, estimated time (quantitative) and client type (categorical) were identified as the variables that affect the occurrence of delays in a BMS project. But to conduct a linear discriminant analysis, only quantitative predictors are considered. The significant quantitative variables being, the number of points and the estimated time for a BMS project. In this case, the response variable indicated the two groups named as delayed and not delayed. Also, by observing the matrix plot (figure 2) it was assumed that the points and estimated time have low multicollinearity and the calculated correlation between points and estimated time was 0.455.

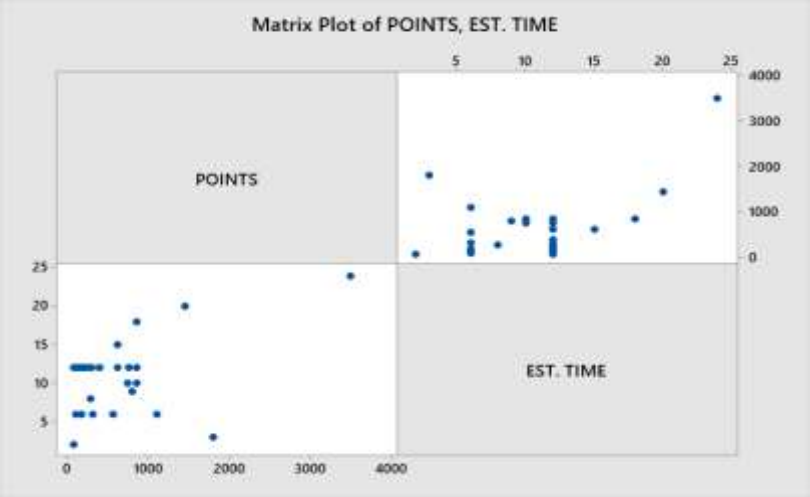

Fig -2: Matrix plot for multicollinearity of points and estimated time

The p-value of Bartlett's test for equal variances of the number of points between two groups delayed and not delayed was 1.000. Since p-value >0.05, there is not enough evidence to reject the null hypothesis, i.e. all variances of points are equal between delayed and not delayed groups with 95% confidence. Also, the p-value of Bartlett's test for equal variances of estimated time between two groups delayed and not delayed BMS project was 0.836. Since p-value > 0.05, there is not enough evidence to reject the null hypothesis; all variances of estimated time are equal between delayed and not delayed groups with 95% confidence. Hence, both the number of points and estimated time had equal variances in both delayed and not delayed BMS project groups.

Table -1: Summary of Classification Table for Linear Discriminant Models

Fold	Sensitivity	Specificity	Accuracy
1st	0	0.4	0.25
2nd	1	0.5	0.625
3rd	1	0.75	0.8571
4th	1	0.4	0.625

With the satisfaction of these assumptions, the linear discriminant functions were constructed for every 4 folds and validated by using a classification table. In the classification table (table 1), sensitivity reveals the proportion of projects which were correctly classified as delayed projects while the specificity reveals the proportion of projects which were correctly classified as not delayed. According to the classification table, the best overall accuracy occurred in 3rd fold validation and it was 0.8571. The sensitivity of that model was 1, while the specificity of the model = 0.75. Hence the best discriminant model to predict the occurrence of delays in the BMS project was:

LDS (Yes) = -2.8106 + 0.5315* Estimated time
LDS (No) = -2.4141 – 0.0007* Points + 0.5287* Estimated time

3.3 Binary Logistic Regression Analysis

The binary logistic regression model is the second model implemented to predict whether future BMS projects can be delayed or not delayed. The advantage of constructing the binary logistic regression model was both categorical and quantitative variables were considered to predict the occurrence of delays. Hence the binary logistic regression models were constructed for each fold considering the number of points, estimated time for a BMS project as well as client type as predictors. To construct the model, the reference levels were selected as shown in table 2.

Table -2: Reference Levels of Binary Logistic Regression Models

Variable	Reference level	Reason for taking reference level
Delay	Yes	The implementation of the model is interested to identify the occurrence of delays in BMS projects.
Client Type	Government	The occurrence of delays in BMS projects for government is high.

According to the goodness of fit, in the 1st, 2nd and 4th fold, the p-value of the Hosmer-Lemeshow test is greater than 0.05, i.e. there isn't enough evidence to conclude that the model does not fit the data. Therefore, the binary logistic regression model for those folds fits well with the considered data with 95% confidence. But in the 3rd fold, the p-value of the Hosmer-Lemeshow test was 0.029. Since the p-value < 0.05, there isn't enough evidence to conclude that the model does not fit the data. Therefore, the binary logistic regression model for 3rd fold did not fit well to the considered data with 95% confidence.

Table -3: Summary of Validation in Binary Logistic Regression Models

Fold	Sensitivity	Specificity	Accuracy	AIC	P-Value
1st	0.6	1	0.75	28.65	0.818
2nd	0.75	0.5	0.625	32.33	0.191
3rd	-	-	-	35.71	0.029
4th	1	0.5	0.75	32.15	0.225

According to table 3, both 1st fold validation and 4th fold validation had the same accuracy 0.75. Akaike Information Criterion (AIC) is a measure of the relative quality of a model that accounts for fit and the number of terms in the model, and it is used to compare different models. Hence by considering the AIC values of 1st and 4th binary logistic models implemented, concluded the 1st model as the best binary logistic model to predict the occurrence of delays in BMS project as it had the least AIC value (28.65) corresponding to the AIC value of 4th model (32.15). Sensitivity and the specificity of that model were 0.6 and 1 respectively. Therefore, the best binary logistic model to predict the occurrence of delays in the BMS project was:

Government**:** Y' = 12**.**43 + 0**.**004979 Points - 0**.**07640 Estimated time
Private**:** Y' = -0**.**6228 + 0**.**004979 Points - 0**.**07640 Estimated time

The primary difference between these two models is the predictors. To construct the discriminant function, only quantitative variables should be considered as predictors. But by the descriptive analysis, it is identified that the categorical variable client type also affects the occurrence of delays in BMS projects as well as quantitative variables number of points and estimated time. Hence the predictions of discriminant function are based on points and estimated time of the BMS project. In contrast, the predictions of the binary logistic function are based on the number of points, estimated time and client type. Linear discriminant function is much similar to the binary logistic model. But there are many assumptions and restrictions in the discriminant analysis compared to binary logistic regression. Such one assumption of linear discriminant analysis is that the normality of predictors for each group. It was not identified that points and estimated time of BMS projects are normally distributed with a 0.05 significance level for each group by Anderson Darling test for normality. Still, by assuming the points and estimated time is normally distributed for each group, the linear discriminant function was implemented. However, according to Alayande & Adekunle, when discriminant analysis' assumptions are met, it is more powerful than the binary logistic regression [1]. Because of the low multicollinearity, the predictions of the linear discriminant function could be better than the predictions of the binary logistic model. The researchers who are interested in this study can conduct the study with a sufficient sample size to gain similar, more accurate results to predict the occurrence of delays in BMS projects.

4. LIMITATIONS

The limitations of this study are as follows; the limited size of sample data, non-accessibility to senior management for obtaining views and only applicable for BMS projects that outsource NVQ qualified technicians**.**

5. CONCLUSION

The number of points, estimated time and the client type of a BMS project were identified as the potential parameters that affect delay occurrence. By considering the accuracy of the implemented best linear discriminant function (0.8571) and the best binary logistic model (0.75), it can be concluded that the linear discriminant function is better equipped to predict delay occurrence in a BMS project than a binary logistic function.

6. REFERENCES

[1]. Alayande, S & Adekunle, B 2015, *'An overview and application of discriminant analysis in data analysis'*, IOSL Journal of Mathematics, *vol. 11, iss. 1, pp. 12-15, viewed 7 February 2020, http://www.iosrjournals.org/iosr-jm/papers/Vol11issue1/Version-5/B011151215.pdf*

[2]. Glen, S 2016, *Hosmer-Lemeshow Test: Definition*, Statistics How To, viewed 17 February 2020, https://www.statisticshowto.datasciencecentral.co/hosmer-lemeshow-test/

[3]. Hosmer, DW & Lemeshow, S 2013, *Applied Logistic Regression*, 3rd edn, E-book, John Wiley & Sons, Inc., Toronto, available at https://media.wiley.com/product_data/excerpt/72/04705824/0470582472-45.pdf

[4]. James, G, Witten, D, Hastie, T & Tibshirani 2014, *An Introduction to Statistical Learning with Applications in R*, E-book, Springer, Berlin, available at http://faculty.marshall.usc.edu/garethjames/ISL

[5]. Kassambara, A 2018, 'Multicollinearity Essentials and VIF in R', *Regression Model Diagnostics*, viewed on 1 February 2020, http://www.sthda.com/english/articles/39-regressionmodel-diagnostics/160-multicollinearity-essentials-and-vif-in-r/

[6]. National Institute of Standards and Technology International SEMATECH 2003, *Engineering Statistical Handbook*, E-book, https://www.itl.nist.gov/div898/hand book/

[7]. Olson, DL & Delen, D 2008, *Advanced Data Mining Techniques*, E-book, Springer, Berlin, available at https://books.google.lk/books?id=2vbLZEn8uUC&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

[8]. Rana, R & Singhai, R 2015, 'Chi-square Test and its Applications in Hypothesis Testing', *Journal of the Practice of Cardiovascular Science*, vol. 1, iss. 1, pp. 69-71, viewed 1 February 2020, http://www.j-pcs.org/temp/JPractCardiovascSci1169-3 448838_0 93448.pdf

[9]. Trochim, WMK 2004, *Research Methods Knowledge Base*, 2nd edn, E-book, Atomic Dog Publishing, Cincinnati, OH, available at http://indexof.co.uk/Various/Social%20Research%20Methods%20Knowledge%20Base.pdf

[10]. Zaiontz, C 2018, *Classification Table*, Real Statistics using Excel, viewed 16 February 2020, http://www.realstatistics.com/logisticregression/classification-table/