Analysis of Port Fare Increases On Container Yard Services Using Logistic Regression Model

S Zuhri1*, P D Sentia1, N A Lubis1 S D Permai2
1Industrial Engineering Department, Universitas Syiah Kuala, Banda Aceh, 23111, Indonesia
2Statistics Department, School of Computer Science, Bina Nusantara University, Indonesia
*Email: sarika.zuhri@unsyiah.ac.id

Abstract. The port company used in this research is the only port where container ship contained in Aceh and has been operating for more than 1 year. The problem with this was the large buildup of containers with long periods of time in the cultivation field that would cause losses. Losses due to the duration of the buildup are influenced by several factors, such as the yard of ratio and the length of time the container is reviewed based on the large number of containers in the container yard. For payment of the first five days of service, the container accumulation fare shall be borne by container owner company as stakeholder of port company. After more than five days, the customer will be charged the container accumulation rate calculated per box without any time consideration. This causes the customers don’t take the goods immediately, therefore it was necessary to do research on the analysis of port fare increase in port company in Aceh using logistic regression model as the basic of decision making. Based on the prediction of logistic regression model using software R3.4.3, there was a loss cost in the container yard, therefore it was necessary to take decision of the application of the port fare increase on container yard service. Based on the results of these predictions, the resulting model was a good model with an accuracy of 74.8%. The factor that affect of the loss cost was the length of time the container in the container yard period II and III.

1. Introduction
Indonesia is an archipelago with a strategic location so that, Indonesia needs a good logistics system. Ports have an important role as one of the distribution chain of goods, which is a logistics process. Based on the logistics performance index of Indonesia in 2016 decreased by 10 ratings ie the order of 63 of the 160 countries [1]. To handled this, presidential of Indonesia plans decreasing in logistic costs by 20 to 25%.

A way the government is lowering logistics costs by simplifying the examination by pressing the number of time loading and unloading and set fines of up to 900 percent of port rates at Tanjong Priok Port as the reference to the three ports such as Tanjong Perak, Belawan and Soekarno Hatta port. The presence of the port fare increase plan make some pros and cons.

Port company used in this research is one of 24 ports that includes in the program sea motorway. With the marine highway program is expected to improve the performance of logistics and lowering the loading and unloading time so, no more excessive accumulation of goods for a long time import. The accumulation of goods is too long portis one of the things that should be minimized. In addition to bringing the ineffectiveness also cause of loss cost for itself. Losses incurred due to the length of the buildup allegedly influenced by several factors: the yard of the ratio and the length of time the container that is period I for 1 to 5 days, the period II for 6 to 9 days, and the period III for the great time equal to 10 days that reviewed by the large number of containers in the container yard. The load capacity can be stacked in the container yard of Port company is 500 (Twenty-Foot Equivalent Units).

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Based on current conditions, for payment of container yard service usage after more than five days, it is not covered by container owner company as stakeholder of port company. So, the customer will be charged a container stacking which are calculated per box without any consideration of the time. Because there is no consideration of time in the stacking fare, it makes the customer does not take the goods immediately. To minimize the inefficiencies and losses because of stacking container are too long, the necessary basis for decision making in determining the rate increase on services which consider the accumulation of time in port company in Aceh.

2. Methodology and Conceptual Framework
Port company in Aceh has problem that can be seen based on survey data. Survey data is the data from port company that is used in this research. Survey data consist of two part such as dependent and independent variable. Each data of survey data consists of 12 month that started from May 2017 to April 2018.

Making the decision in a firm is very important thing. Logistic regression is one of tools as part of regression that can be used for decision making. The difference between logistic regression and linear regression is a dependent variable of the linear regression is continuous and dependent variable in logistic regression is dichotomous, It can be seen in a regression analysis when variable dependent is the dichotomy so a regression equation must be formulated to be limited between 0 and 1. The distribution is binomial distribution, not normal. It describe the distribution of the error and will be a statistical distribution is the basis of its analysis, and principles used in the linear regression will also be a guide in logistic regression [2].

Based on previous research logistic regression model is a model that can be used in solving the problem of decision making [3]. Logistic regression model is also an effective model to detect and predict an issue [4,5,6,7] and a logistic regression model can also be used in evaluating the risk that logistic regression is a model that is attractive and effective techniques [8,9,10].

The variables used in this study to build a logistic regression model that is

| Notation | Name               | Definition                                                                 | Information       |
|----------|--------------------|---------------------------------------------------------------------------|-------------------|
| Y        | There is or there is no loss cost | Variable presence or absence of costs or loss arising from the container stacking in the container yard. | 0= No loss 1= loss |
| X1       | Yard of Ratio      | Comparison of stacking yard capacity consumed with overall capacity container yard | Continuous        |
| X2       | Number of Containers | Number of existing container Stacking in the container yard | Continuous        |
| X3       | The length of time the container | The length of time of the container that’s located in the container yard | 0= Period I 1= Period II 2= Period III |

The logistical function can ensure that whatever is needed, there will always be between 0 and 1 which is the reason that the logistic model is available to estimate the opportunity [10]. The specific form of the logistic regression model or regression model is [2]

\[
\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}
\] (1)
Making more easier to estimate regression parameters, \( \pi(x) \) is transformed using logit transformation. Logit transformation in the logistic regression model is

\[
g(x) = \beta_0 + \beta_1 x
\]

Where,
\[
\pi(x) : \text{Probability of occurrence} \\
g(x) : \text{Logit estimation value} \\
\beta_0, \beta_1 : \text{Coefficient values for the constant variables obtained using the estimation method} \\
x : \text{Independent variable}
\]

what can be seen in a regression analysis is a regression equation must be formulated to be limited between 0 and 1, binomial distribution not normal, describes the distribution of errors and will be a statistical distribution which is the basis of the analysis, the principle used in regression linear will also be a guide in logistic regression. Here is the value of the logistic regression model if the independent variable is dhyctomous

| Table 2. Value of Logistic Regression Model If Independent Variable is Dhyctomous |
|-------------------------------|------------------|------------------|
| **Dependent variable (Y)** | **Independent variable (X)** |                      |
|                               | **X=1** | **X=0** |
| **Y=1** | \[\pi(1) = \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}}\] | \[\pi(0) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}\] |
| **Y=0** | \[1 - \pi(1) = \frac{1}{1 + e^{\beta_0 + \beta_1}}\] | \[1 - \pi(0) = \frac{1}{1 + e^{\beta_0}}\] |
| **Total** | \[1,0\] | \[1,0\] |

Logistic regression significance test consists of two stages: test the significance of the model parameters together and test the significance of the model parameters separately. A likelihood ratio goodness of fit (G) test statistic is a test of the significance of the model parameters together which is used to examine the role of predictor variables in the model together. The formula for the G likelihood ratio test is

\[
G = -2 \ln \left( \frac{L_0}{L_p} \right)
\]

Where,
\[
L_0 : \text{Likelihood with no independent variable} \\
L_p : \text{Likelihood with p independent variable}
\]

Another test phase is Wald test which is obtained by comparing the estimation of maximum likelihood from the slope parameter to the estimated standard of errors. Wald test is an individual test on the significance of the model parameters defined by

\[
W = \left( \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \right)^2
\]

Where,
\[
\hat{\beta}_i : \text{Estimator of } \hat{\beta}_i \\
SE(\hat{\beta}_i) : \text{Default error estimator of } \hat{\beta}_i
\]
One of the classification accuracy tests used for categorical data is apparent error rate (APER). A good model is a model that produces a slight misclassification. APER is defined as a comparison of observations in a sample of training data that is incorrect by the sample classification function where, APER is a value that is used to see opportunities for errors in classifying objects. Apparent error rate (APER) can be easily calculated from the confusion matrix, which shows the actual group membership versus the predicted one. The form of the confusion matrix is:

Table 3. Confusion Matrix

| Actual member | Predicted member | n1C | n1M | n2C | n2M |
|---------------|-----------------|-----|-----|-----|-----|
| n1            | n2              | n1C | n1M | n2C | n2M |

Where,
- \( n_{1C} \) = Number of data \( n_1 \) correctly classified as data \( n_1 \)
- \( n_{1M} \) = Number of data \( n_1 \) incorrectly classified as data \( n_1 \)
- \( n_{2C} \) = Number of item \( n_2 \) correctly classified
- \( n_{2M} \) = Number of item \( n_2 \) incorrectly classified

The formula to calculate APER is

\[
APER = \frac{n_{1M} + n_{2M}}{n_1 + n_2}
\]

(5)

3. Result and Discussion

The following are the results of modeling and parameter estimation using logistic regression models. Modeling is done using one dependent variable and three independent variables.

Table 4. Output of logistic regression model

| Variable | \( \beta_0 \) | \( X_1 \) | \( X_2 \) | \( X_{32} \) | \( X_{33} \) |
|----------|-------------|--------|--------|--------|--------|
| Yard of Ratio \((X_1)\) | -3.491 | 0.004 | -0.003 | 4.386 | 3.230 |

Based on Table 2 model created is:

\( g(x) = -3.491 + 0.004X_1 - 0.003X_2 + 4.386X_{32} + 3.230X_{33} \)

Based on the logistic regression model, the output summary conclusions model which can be seen in value of p-value of each variable.

Table 5. Result of logistic regression model test

| Variable | p-value | Information |
|----------|---------|-------------|
| Yard of Ratio \((X_1)\) | 0.659 | Not significant |
| Number of Container \((X_2)\) | 0.538 | Not significant |
| The length of time the container period II \((X_3)\) | 4.22e-12 | Significant |
| The length of time the container period III \((X_3)\) | 1.40e-07 | Significant |
Based on Table 5 it can be seen that the independent variables that influence whether or not the loss cost is variable $X_3$, namely the length of the time period II and III. It can be seen from the p-value $\leq \alpha$ that is $4.22\times10^{-12}$ dan $1.40\times10^{-07}$. Variables that do not affect whether or not the loss is a $X_1$ variable that is yard of ratio and $X_2$ that is the number of containers in the container yard. It can be seen from the value of p-value that is greater than $\alpha$. Here's a likelihood ratio test output which is a significance test parameters simultaneously

| Table 6. Output of ratio likelihood test in model |
|-----------------------------------------------|
| LogLik | Chisq  | Pr(>Chisq) |
|--------|--------|------------|
| -174.26|        |            |
| -233.24| 117.96 | <2.2e-16   |

The hypothesis are:
$H_0$ = No influence of variables - independent variables to whether or not the loss cost
$H_1$ = At least there is one independent variable which affects whether or not loss cost

To calculate the value of G that is by comparing the value of G obtained in the Table

$G = -2 (\text{-233.24-(-174.26)})$

$G = 117.96$

Rated G obtained in the amount of 117.96 that will be compared with the value of chi-square Table with df = 4(\(X^2_{0.05,4}\) is 9.488. If \(G \geq \text{Chi square Table}\) then \(H_0\) rejected and \(H_1\) accepted. Based on the results of the likelihood ratio test \(G \geq \text{Chi square Table value}\) is equal to 117.96 ≥ 16.92. This means that, at least there is one independent variable which affects the loss cost in the container yard.

Wald test is used to determine the independent variables that affect the model. \(H_0\) is rejected if P(Wald's Test) ≤ $\alpha$. Hypothesis in Wald test are :

$H_0$ = Independent variable $k$ on \(g(x)\) is not significant effected on the model.
$H_1$ = Independent variable $k$ on \(g(x)\) is significant effected on the model.

| Table 7. Output of Wald test in model |
|--------------------------------------|
| Variabel                             | P(Wald’s Test) |
| Yard of Ratio ($X_1$)                | 0.659           |
| Number of Container ($X_2$)          | 0.538           |
| The length of time the container period II ($X_3$) | <0.001 |
| The length of time the container period III ($X_3$) | <0.001 |

Based on the Wald test output in Table 7, namely yard of $X_1$ and $X_2$ ratio is the number of containers is an independent variable that does not affect the model because, the value of P (Wald's Test) ≥ $\alpha$ are 0.659 and 0.538. $X_3$ is the length of time the buildup of future category II and III period are independent variables that affect the model, because nilai P (Wald's Test) ≤ $\alpha$ is 0.001.

Fitting model test is performed to evaluate the model fit with data. Model fit test that used was Hosmer and Lemeshow test. \(H_0\) is rejected if the p-value ≤ $\alpha$. Hypotheses used are:

$H_0$ = Logistic regression model fit to the data
$H_1$ = Logistic regression model did not fit to the data
Here is the output from the model fit test using Horsmer and Lameshow test.

**Table 8. Output of Horsmer and Lameshow test in model**

| Horsmer and Lameshow | P-value  |
|-----------------------|----------|
|                       | ≤2.2e×10^{-16} |

Based on Table 8, the value of $p$-value $\leq \alpha$, that is $2.2e\times10^{-16} \leq 0.05$. It means that the logistic regression model formed is not fit to the data. This happens because of from the five independent variables were used only two variables that affect whether or not the cost of loss.

The means used to interpret the parameter is to look at the value of the odds ratio. Odds ratio results of the data are shown in Table 9

**Table 9. Result of odds ratio test in model**

| Variabel                          | Odds Ratio |
|-----------------------------------|------------|
| Yard of Ratio (X1)                | 1,004      |
| Number of Containers (X2)         | 0.997      |
| The length of time the container Period II (X3) | 80.35     |
| The length of time the container Period III (X3) | 25.27     |

According to the Table 9, can be interpreted that each increase of one unit of a variable length of time the container category period II, the opportunity cost of the loss due to accumulation of containers in port company is 80.35 times compared to category period III.

Evaluation of the regression model described by pseudo $R^2$, but because of the logistic regression using R3.4.3 there are not pseudo $R^2$ value, then the index McFadde can be used to assess the percentage of independent variables able to explain the dependent variable of the logistic regression model that has been formed. McFadden output results can be seen in Table 10.

**Table 10. Result of evaluation of model**

| $G^2$       | McFadden | $R^2$ML |
|-------------|----------|---------|
| 117,961     | 0.253    | 0.276   |

$R^2$ McFadden value = 0.253 it means the independent variable of the existing logistic regression model can explain the variability of the dependent variables of 25.3%, while the remaining 74.7% is explained by other variables not included in the model.

How to read the results predicted in Figure 1 is to compare the predicted results using a logistic regression model with dependent variable survey data.
Based on the predicted results in Figure 1, there is the loss cost by the symbol 1. Cost loss arising on the predicted outcome means that the losses incurred due to the length of the stacking containers in container yard of port company so, the decision to port fare increase on container yard services is the best decision.

Calculate the classification accuracy of the model is done by calculating the value of the apparent error rate (APER). Confusion matrix for survey data can be seen in Table 11.

The results of the model's accuracy survey data is:

\[
1-\text{APER} = 1 - \frac{27+65}{58+27+65+215} = 1 - \frac{92}{344} = 0.748 
\]

Based on the results of the accuracy survey data, the classification accuracy of determining the presence or absence of the loss cost in determining the need for a port fare increase is 74.8%. This shows that the model is already well in the presence or absence of classifying the loss cost in determining the need for a decision on the port fare increase in container yard services of port company.

4. Conclusions

The conclusions of this research are \( g(x) = -3.491 + 0.004X_1 + 0.003X_2 + 4.386X_3^2 + 3.230X_3^3 \). The factors affecting the rise of the loss cost in the container yard of port company are the duration of the container period II and period III. Based on the prediction results using logistic regression there is the loss cost. Therefore it is necessary to take the decision of the port fare increase in container yard service. So the resulting model is a good model with an accuracy of 74.8%.
Acknowledgments
The authors are thankful to the industrial engineering department of Syiah Kuala University and Bina Nusantara University for having funded this study.

References
[1] Arvis, J. Mustra, M. Panzer, J. Ojala, L. Naula, T. (2016) ‘Connecting To Compete: Trade Logistics In The Global Economy’, World Bank, pp. 1–76.
[2] Hosmer, D. W. And Lemeshow, S. (2000) ‘Applied Logistic Regression’. John Wiley Series In Probability And Statistics, Inc, New York.
[3] Mehrjoo, S. and Bashiri, M. (2013) ‘An Application Of Principal Component Analysis And Logistic Regression To Facilitate Production Scheduling Decision Support System: An Automotive Industry Case’. Journal of Industrial Engineering., pp. 2-12
[4] Chandrasekaran, R. (2012) ‘Application Of Logistic Regression To Predict Over Target Baseline Of Software Projects’, 44(April), pp. 1–6.
[5] Agarwal, S., Kachroo, P. And Regentova, E. (2016) ‘A Hybrid Model Using Logistic Regression And Wavelet Transformation To Detect Traffic Incidents’, IATSS Research. International Association Of Traffic And Safety Sciences, 40(1), pp. 56–63.
[6] Christine, L. Aderson, A. Karoui, H. Podgorski, J. Yacouba, H. (2018) ‘Catena Prediction Model For Cyanide Soil Pollution In Artisanal Gold Mining Area By Using Logistic Regression’, Catena. Elsevier, 162 (November 2017), pp. 40–50.
[7] Salas-Eljatib, C. Fuentez-Ramirez, Andres. Gregoire, Timothy. Altamirano, Adison. Yaitul, Valeska. (2018) ‘A Study On The Effects Of Unbalanced Data When Fitting Logistic Regression Models In Ecology’, 85(August 2017), pp. 502–508.
[8] Goh, A. T. C. Zhang, Yanmei. Zhang, Runhong. Zhang, Wengang. Xiao, Yang. (2017) ‘Evaluating Stability Of Underground Entry-Type Excavations Using Multivariate Adaptive Regression Splines And Logistic Regression’, Tunnelling And Underground Space Technology. Elsevier, 70(August), Pp. 148–154.
[9] Morotti, S. and Grandi, E. (2017) ‘Methodx Logistic Regression Analysis Of Populations Of Electrophysiological Models To Assess Proarrhythmic Risk’, Methods X. Elsevier B.V., 4, pp. 25–34.
[10] Zhu, Z. Lin, C. Zhang, X. Wang, K. Xie, J. Wei, S. (2018) ‘Evaluation Of Geological Risk And Hydrocarbon Favorability Using Logistic Regression Model With Case Study’, Marine And Petroleum Geology. Elsevier, 92(February), Pp. 65–77.