Journal of Actuarial, Finance and Risk Management (JAFRM)
Vol. 1(1), June, 2022.

Logistic Regression Analysis of Demographic and Vehicle Condition for Purchasing Vehicle Insurance

Gabriel Azhar[1], Muhammad Cahirul Rahman[2], Rosyid Nur Salam[3], Fauziah Nur Fahirah Sudding[4*]

[1,2,3,4] *Study Program of Actuarial Science, School of Business, President University, 17550, Indonesia*
**Corresponding author: fauziahfahirah@president.ac.id*

Abstract— Insurance is a contract, represented by a policy, in which an individual or entity receives financial protection or reimbursement against losses from an insurance company. Insurance policies are used to hedge against the risk of financial losses, both big and small, that may result from damage to the insured or her property, or from liability for damage or injury caused to a third party. Building a model to predict whether a customer would be interested in Vehicle Insurance is extremely helpful for the company because it can then accordingly plan its communication strategy to reach out to those customers and optimize its business model and revenue. In this research, we use the secondary data that collected in India in 2020, which analyzes vehicle condition, demographics, and owning a driver's license on vehicle insurance buying interest. The method used in this research is the Logistic Regression, the response variable is the Response (of buying vehicle insurance interest), and the independent variables are Gender, Driving License, Previously Insured, Vehicle Age, and Vehicle Damage. The result of this research showed that the Previously Insured, Vehicle Age, and Vehicle Damage have a correlation to the Response.

Keywords— Demographic; Logistics Regression; Purchase; Vehicle Insurance.

I. INTRODUCTION

Traffic accidents have always been a global concern and according to World Health Organization (WHO), they are estimated to be eight leading causes of death for all group of age. The incident occur for various reasons. Most of the common causes are human mistakes, vehicle conditions, and bad weather. Each year, there are approximately 1.3 million people die due to this incident. This incident not only take lives but also cause financial loss to the individuals. Hence, it is very important to anticipate this unexpected event in the future by considering vehicle or auto insurance as financial protection.

Vehicle insurance is a legal contract between the insurance company and policyholder to provide compensation to the insured due to unexpected losses to the vehicle against traffic injuries or any kind of damages. It is the best form of security to prevent significant payment as a result of calamities. Auto insurance does not only cover full or partial loss sustained by the insured vehicle, but also offers coverage for third-party liabilities. Due to economic issues, getting insurance is used to be not a priority for most people, especially in developing countries. However, auto insurance is now being recognized and even obligated in some countries to drive legally. Therefore, building a model to project the number of potential customers is beneficial and informative to the company since it can then accordingly plan its communication strategy and optimize its business model and revenue.

Some researchers have developed some studies to analyze customer behavior. For example, Guillen, M., et all. (2002) used log-linear regression to analyze customer loyalty in insurance industry. The result of this study indicates customers having motor insurance have much higher risk of canceling one of their policies and the higher risk period to cancel are between 2 and 4 years since their first policy issue. In a separate study, Dragos, C. M., and Dragos, S. L. (2017) also analyze customer behavior in auto insurance but using a different method. The researchers used discrete choice models to define two behavioral factors when customers decide to purchase new motor insurance policy. Those two factors are the decision to purchase the insurance or not and the motive behind the non-subscription of the insurance policy. The results show that an individual willingness and a lower education level factors increase the likelihood of not buying a motor insurance policy. On the other hand, factors such as the willingness not to purchase insurance, car annual mileage, and the ratio between estimated car price and customer wage are statistically significant in terms of motivation to not buy an insurance policy. An, S. H., Yeo, S. H., and Kang, Minsoo. (2021) compared two methods to build a model for purchasing automobile insurance. The authors use two-class logistic regression and two-class boosted decision tree. The results show that these two models can be used to design a new model with high accuracy to project customers in the future even though the number of potential insured does not significantly increase. Another study about customer behavior is Duan, Zhengmin.,et all. (2018) use logistic regression to auto burden index to design a new rate-making model for auto insurance. The authors use data from Chinese insurance company and the results show that the auto burden index has good-fitted to rate-making model. However, due to the limitation of data source, the authors found that the coefficient of this factor is not really significant in the model.

According to the studies above, the authors conclude that log-linear regression is one of the statistical methods that will be good-fitted to be used for building a model in the insurance industry and it encourages authors to study customer behavior in subscribing insurance policy using this method. Variables such as response, gender, driving license, previously insured, vehicle age, and vehicle damage are chosen to be used in this paper, and the type of data used is secondary data collected in India in 2020. The aim of this study is to analyze the association between the variables, so it can be used to determine the best-fitted model for projecting a potential number of customers for vehicle insurance using log-linear regression.

II. LITERATURE REVIEW

This regression is statistical modeling for response variables in binary data, the outcome will be success and failure (Nugraha, 2014). Given Y (the response variable) and X (the explanatory variable), assume the relationship of $Y = \beta_0 + \beta_1 X$, here $E(Y|X)$ is a random variable.

A. Binomial Distribution and Logistic Regression

Let y_i become response variable which has a binary data (value 1 or 0)

$$y_i = 1 \text{ if it is success}$$

$$y_i = 0 \text{ if it is failure,}$$

With y_i is a realization of a random variable from Y_i. the probability of Y is

$$P(Y_i = 1) = \pi i \text{ and } P(Y_i = 1) = (1 - \pi_i) \tag{1}$$

Yi has a Bernoulli Distribution with parameter π_i, then we can write

$$P(Y_i = y_i) = \pi_i^{yi}(1 - \pi_i)^{1-yi} \tag{2}$$

For y_i= 0, variance of y_i will be

$$E(Y_i) = \mu_i = \pi_i \text{ and } Var(Y_i) = \pi_i(1 - \pi_i) \tag{3}$$

Parameter π_i is based on X, which is

$$\pi_i = \pi_i(X_i)$$

Since variable of X_i is independent (predictor), we can assume that the predictor doesn't affect the mean and variance itself, then this condition doesn't fit into a binary data response. Example, let n_i~ a sample size in group *i*, and y_i ~ a value of success in group *i*, then

$$P(Y_i = y_i) = \binom{n_i}{y_i} \pi_i (1 - \pi_i)^{n_i - y_i} \tag{4}$$

and

$$E(Y_i) = \mu_i = n_i \pi_i \text{ and } Var(Y_i) = n_i \pi_i (1 - \pi) \tag{5}$$

Where, if $n_i = 1$, means the binomial distribution becomes Bernoulli distribution.

B. Logistic Regression Model with a Single Independent Variable

Recall that Y (the response variable) and X (the explanatory variable), become Y (the response variable binary) and X (predictor variable). Let $\pi(x)$ become a probability of success, then $\pi(x)$ will be a parameter in the Binomial distribution. Let $Y = 1$ and X is the dependent variable, so the regression will be

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{6}$$

and for the probability of logit has a linear function, such as:

$$\log\ it\ [\pi(x)] = \log(\frac{\pi(x)}{1 - \pi(x)}) = \beta_0 + \beta_1 x \tag{7}$$

C. Logistics Regression Model with Many Independent Variables

Logistics regression is a regression that uses two different values to express its responsiveness (Y), usually the values 0 (fail) and 1 (success) are used. The distribution function used is the logistic distribution with the notation π(X) to express the conditional mean of Y if given the covariate vector X = $(x_1, x_2, \ldots\ldots x_p)^T$.
The logistic regression model is:

$$\pi(x) = \frac{\exp(X^T\beta)}{1+\exp(X^T\beta)} \tag{8}$$

with $X^T\beta = \beta_0 + x_1\beta_1 + \ldots + x_p \ \beta_p$
$\beta = (\beta_0, \beta_1, \ldots\ldots, \beta_p)^T$ is a parameter vector.

Defined as a *logit* π(X) transformation, namely:

$$g(x) = \log\frac{\pi(x)}{1-\pi(x)} = X^T\beta \tag{9}$$

So that g(X) is linear in the β parameter.

D. Maximum Likelihood Estimator for Logistic regression

To determine the regression model, the value of β is estimated first using the maximum likelihood method. Log-likelihood function:

$$\log L(\beta) = \sum_{i=1}^{n}\{y_i \log(\pi_i) + (n_i - y_i)\log(1-\pi_i)\} \tag{10}$$

The formula above is derived from the following formula

$$P(Y_i = y_i) = \pi_i^{yi}(1-\pi_i)^{1-yi} \tag{11}$$

From this log-likelihood function, the first and second derivatives are sought. The parameter estimate β is the value of βwhich maximizes the log-likelihood function on the sample data (X, Y).

$$\frac{\partial \log L(\beta)}{\partial\beta} = 0 \tag{12}$$

and

$$H(\beta) = \frac{\partial^2 \log L(\beta)}{\partial\beta\ \partial\beta^T} \tag{13}$$

Matrix H (β)is called the Hessian matrix which is a negative definite matrix.
Based on the first derivative condition, the β parameter estimation using the maximum likelihood method is:

$$\sum_{i=1}^{n}[yi - \pi(Xi)] = 0 \ \text{ and } \ \sum_{i=1}^{n} x_{ij}[yi - \pi(Xi)] = 0 \tag{14}$$

For j = 0,1, ..., p

The second derivative of the log-likelihood function for all parameters is called the Hessian (H) matrix which has elements:

$$\frac{\partial^2 \log L(\beta)}{\partial\beta_j^{\ 2}} = -\sum_{i=1}^{n} x_{ij}^{\ 2}\pi_i(1-\pi_i) \tag{15}$$

and

$$\frac{\partial^2 \log L(\beta)}{\partial\beta_j\ \partial\beta_\mu} = -\sum_{i=1}^{n} x_{ij}x_{i\mu}\pi_i(1-\pi_i) \tag{16}$$

E. Logistics Regression Inference

We have learned how logistic regression helps to illustrate the effect of predictor variables and binary response. Parameters in the logistic model can be estimated using the Maximum Likelihood Estimation Method (MLE). This section will discuss the way estimator properties are being used to infer the parameter [1].

1) *Confidence Interval*

If there is a huge number of samples, the interval confidence of β_j in logistic regression model:

$$\text{Logit } [\pi(x)] = \beta_0 + x_1\beta_1 + \ldots + x_p\beta_p$$

with coefficient interval

$$\hat{\beta}_j \pm z_{\alpha/2}\sqrt{Var(\hat{\beta}_j)} \tag{17}$$

where $\hat{\beta_j}$ is the MLE of β and j = 0,1, ...,p.

2) *Significance Test*

To test the hypothesis H0:β_j= 0, in large samples can be used test statistics

$$z = \frac{\hat{\beta}_j}{\sqrt{Var(\hat{\beta}_j)}} \tag{18}$$

The z statistic has a standard normal distribution.

$$z^2 = \left(\frac{\hat{\beta}_j}{\sqrt{Var(\hat{\beta}_j)}}\right)^2 \tag{19}$$

The z^2statistic has Chi-square distribution with degree of freedom 1. This z statistic is called the Wald test statistic. The likelihood-ratio is more effective compared to the Wald test statistic in the terms of large samples, more trusted, and mostly being used in practice.

3) *Model Fit test*

The measure of deviants can be used to determine whether the model fits the data or not and calculate how much it fits. The deviation statistic (D) measures the mismatch between the observed and expected value in the model.

$$D = -2\sum_{i=1}^{k}\left\{y_i \log\left(\frac{n_i\pi_i}{y_i}\right) + (n_i - y_i)\log\left(\frac{n_i - n_i\pi_i}{n_i - y_i}\right)\right\} \tag{20}$$

or

$$D = 2\sum_{i=1}^{k}\left\{y_i \log\left(\frac{y_i}{n_i\pi_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i\pi_i}\right)\right\} \tag{21}$$

The D statistic has a Chi-square distribution.

In general, variable (X) is separated into two groups

$$X = (X_1, X_2) \text{ and } \beta = (\beta_1 \ \ \beta_2)^T$$

Vector $\beta_1$consists of parameter $p_1$and vector $\beta_2$consists of parameter p_2. Further is hypothesis testing

$$\text{H0:}\beta_2 = 0 \text{ or H0:}\beta_2 \neq 0$$

Null hypothesis (H0) claims that predictor variable in the second group (X_2) does not influence Y response. This is identical with the significance test parameter β_2.

Let D (β_1) defines the deviation value of the model that enter variable $X_1$and D(β) defines the deviation value of the model that enter variable $X = (X_1, X_2)$. Next, the difference between the two deviation values is

$$X^2 = D(\beta_1) - D(\beta) \text{ or } X^2 = -2\log\left(\frac{L(\beta_1)}{L(\beta)}\right) \tag{22}$$

has Chi-square distribution with degree of freedom p_2(for a large sample). Parameter p_2 is the difference in the number of parameters in D (β_1) and D(β).

III. METHODOLOGY

A. *Data Preparation*

This research data is the secondary data taken from Anmol Kumar's Health Insurance Cross-Sell Prediction. Response data for this research is going to be whether customers from past years are interested in buying vehicle insurance.

Table 1 is a slice of the data that will be used later in the analysis. Where it shows the response about Gender, Driving License, Previously Insured, Vehicle Age, and Vehicle Damage and the final response.

TABLE 1
INDEPENDENT AND RESPONSE VARIABLE DATA

Gender	**Driving License**	**Previously Insured**	**Vehicle Age**	**Vehicle Damage**	**Response**
0	1	0	1	1	1
0	1	0	1	0	0
0	1	0	1	1	1
0	1	1	0	0	0
1	1	1	0	0	0
1	1	0	0	1	0
0	1	0	0	1	0
1	1	0	1	1	1
1	1	1	0	0	0
1	1	1	0	0	0
1	1	0	1	1	1
1	1	1	0	0	0
1	1	1	1	0	0
0	1	0	1	1	0
0	1	1	1	0	0
...	....	...	...	...	...

Annotation:

- Gender
  0 = Women; 1 = Men
- Driving license
  0 = Didn't own; 1 = Own
- Previously Insured
  0 = Not insured; 1 = previously insured
- Vehicle Age
  0 = < 1 y e a r; 1 = 1 - 2 y e a r s
- Vehicle Damage
  0 = not damaged; 1 = damaged
- Response
  0 = Didn’t buy insurance; 1 = buy insurance

B. Model p-value Significance Comparison

Model p-value significance comparison is used to see if there is a correlation between the independent variables and the response variable. We have 5 independent variables, which are Gender, Driving License, Previously Insured, Vehicle Age, and Vehicle Damage, and our response variable is Response (of buying vehicle insurance interest). With the help of R software, the result of the p-value for each combination of model is obtained in the Table 2.

Based on the the table , the p-value is obtained for each independent variable, and the hypothesis is as follows:

a. Hypothesis
   H0: There is no correlation between the independent variable to the response variable.
   H1: There is a correlation between the independent variable to the response variable.
b. Significance Level
   $\alpha = 0.05$
c. Critical Area
   Accept H0 if p-value > 0.05
   Reject H0 if p-value < 0.05
d. Decision
   Depends on each model.
e. Conclusion

Each model has a different combination of independent variables, and every independent variable that has p-value greater than 0.05 does not have a correlation to the response variable, and vice versa. If we can see, all models said that Driving License does not have a correlation to the Response. Also, almost all models said that Gender does not have a correlation to the response since only 2 of 40 models have the p-value of Gender smaller than 0.05. Then, we can conclude that the matching model from the p-value significance comparison is the model that has the other three independent variable besides Gender and Driving License, which is Previously Insured, Vehicle Age, and Vehicle Damage. And Bin 3.4.1 is the model.

TABLE 2
MODEL P-VALUE SIGNIFICANCE COMPARISON

Model	Intercept	Gender	Driving License	Previously Insured	Vehicle Age	Vehicle Damage
Bin5	0.000678	0.96	0.471	2.5E-09	2.78E-09	8.89E-10
Bin 4.1	0.0266	0.7022	0.533	< 2e-16	6.68E-10	-
Bin 4.2	0.000507	0.503946	0.555285	9.68E-10	-	9.28E-10
Bin 4.3	5.46E-09	0.884	0.393	-	8E-11	< 2e-16
Bin 4.4	< 2e-16	0.27	-	7.99E-08	1.37E-12	7.82E-13
Bin 4.5	0.944	-	0.957	7.91E-08	1.82E-12	7.92E-13
Bin 3.1.1	0.951	-	0.958	1.16E-14	< 2e-16	-
Bin 3.1.2	< 2e-16	0.368	-	1.11E-14	< 2e-16	-
Bin 3.1.3	0.951	0.547	0.961	-	<2e-16	-
Bin 3.1.4	0.954	0.878	0.958	3.93E-16	-	-
Bin 3.2.1	0.946	-	0.957	2.86E-08	-	6.87E-15
Bin 3.2.2	< 2e-16	0.827	-	2.84E-08	-	7.40E-15
Bin 3.2.3	0.941	0.784	0.958	-	-	<2e-16
Bin 3.2.4	0.226	0.369	0.884	<2e-16	2.42E-12	1.92E-12
Bin 3.3.1	0.000156	-	0.754	-	0.000014600	< 2e-16
Bin 3.3.2	< 2e-16	0.755	-	-	0.000000151	< 2e-16
Bin 3.3.3	0.000953	0.725	0.848	-	-	< 2e-16
Bin 3.3.4	0.00747	0.68016	0.95522	-	< 2e-16	-
Bin 3.4.1	< 2e-16	-	-	5.54E-12	0.00000217	5.24E-11
Bin 3.4.2	< 2e-16	0.755	-	-	0.000000151	< 2e-16
Bin 3.4.3	< 2e-16	0.668	-	1.96E-12	-	3.76E-12
Bin 3.4.4	< 2e-16	0.941	-	< 2e-16	0.000000019	
Bin 3.5.2	0.000156	-	0.754	-	0.000000146	< 2e-16
Bin 3.5.3	0.00592	-	0.80099	1.97E-12	-	3.24E-12
Bin 3.5.4	0.0686	-	0.7783	< 2e-16	1.32E-08	-
Bin 2.1.1	<2e-16	0.369		<2e-16	-	-
Bin 2.1.2	<2e-16	0.799	-	-	<2e-16	-
Bin 2.1.3	<2e-16	0.396	-	-	-	<2e-16
Bin 2.1.4	0.94284	0.00436	0.95016	-	-	-
Bin 2.2.1	0.961	-	0.964	2.96E-16	-	-
Bin 2.2.2	0.936	-	0.949	-	<2e-16	-
Bin 2.2.3	0.951	-	0.964	-	-	<2e-16
Bin 2.3.1	< 2e-16	-	-	5.91E-15	1.67E-12	-
Bin 2.3.2	< 2e-16	-	-	2.15E-08	-	3.50E-12
Bin 2.4.1	< 2e-16	-	-	-	3.72E-10	< 2e-16
Bin 1.1	< 2e-16	0.00494	-	-	-	-
Bin 1.2	0.943	-	0.95	-	-	-
Bin 1.3	< 2e-16	-	-	3.04E-16	-	-
Bin 1.4	< 2e-16	-	-	-	< 2e-16	-
Bin 1.5	< 2e-16	-	-	-	-	< 2e-16

C. Model AIC Comparison

AIC is used when you have multiple model options, determining the quality of each model to be then compared to one another, while finding the best model which has the lower AIC level than another. (Zajic, 2019). Here we will compare the AIC of every model that we have, to filter the better model according to this method. According to the Table 3, we can conclude that model Bin 4.4 and Bin 5 are the models with the lowest AIC.

TABLE 3
MODEL AIC COMPARISON

Model	**AIC**	**Model**	**AIC**	**Model**	**AIC**
Bin 5	2725.74	Bin 3.3.1	2798.14	Bin 2.1.3	2832.10
Bin 4.1	2826.40	Bin 3.3.2	2790.81	Bin 2.1.4	3622.26
Bin 4.2	2763.87	Bin 3.3.3	2834.08	Bin 2.2.1	2893.69
Bin 4.3	2792.80	Bin 3.3.4	3431.82	Bin 2.2.2	3441.20
Bin 4.4	2723.77	Bin 3.4.1	2730.34	Bin 2.2.3	2844.37
Bin 4.5	2732.31	Bin 3.4.2	2851.09	Bin 2.3.1	2831.68
Bin 3.1.1	2833.68	Bin 3.4.3	2761.87	Bin 2.3.2	2773.33
Bin 3.1.2	2824.40	Bin 3.4.4	2824.40	Bin 2.4.1	2796.16
Bin 3.1.3	3431.82	Bin 3.5.2	2798.14	Bin 1.1	3620.91
Bin 3.1.4	2879.99	Bin 3.5.3	2775.32	Bin 1.2	3652.52
Bin 3.2.1	2775.32	Bin 3.5.4	2833.68	Bin 1.3	2891.74
Bin 3.2.2	2761.87	Bin 2.1.1	2878.03	Bin 1.4	3439.33
Bin 3.2.3	2834.08	Bin 2.1.2	3429.94	Bin 1.5	2842.38
Bin 3.2.4	2879.99				

D. Best Model Comparison (Haslem Test)

TABLE 4
HOSLEM TEST

	hstest p value	**hstest x**
Bin 3.4.1	0.98	2.022
Bin 5	0.97	2.36
Bin 4.4	0.97	2.33
qchisq (0.95,8)		15.51

In hoslem test, we compare the p-value of every model that has passed the P-value and AIC comparison. We will then do hoslem test to see whether the p-value and x square passed the test. The p-value is said to be passed from the hoslem test when $p > 0.05$. and x square is said to be passed when hstest x < x square from the table. According to Table 4, we can see that every model passed the test for both p-value and x square value. But model 3.4.1 has the best value because it has the biggest p-value.

IV. RESULTS AND ANALYSIS

After analyzing the models from every combination of variables, we got 3 best models. And by comparing it again using hoslem test as mentioned in 3.3, we got model 3.4.1 consisting of Previously Insured, Vehicle Age, and Vehicle Damaged as the variable. The model itself could be written as

$$Y = -3.21 - 3.66X_1 + 0.55X_2 - 1.6685X_3$$

$$Y = \frac{\exp(-3.21 - 3.66X_1 + 0.55X_2 - 1.6685X_3)}{1 + \exp(-3.21 - 3.66X_1 + 0.55X_2 - 1.6685X_3)} \quad (23)$$

where
X_1 = Previously Insured
X_2 = Vehicle Age
X_3 = Vehicle Damage

This also means that these three variables are the most significant variable for analyzing the probability of someone buying the insurance or not. While the others who doesn't included in the model were not significant enough to be considered.

V. CONCLUSION AND IMPLICATION

According to the analysis above, the authors believe that logistic regression is an effective method to create models to assume the probability of someone deciding to buy vehicle insurance or not. We believe that we have created the best model to be used within the data available and we hope that this model can be applied in insurance companies. Although every company might have different available data and different trends or different types of customers, this model can be used as a comparison or as a consideration. We suggest for future research to be done with different regression or maybe add some tests to compare the result as we got in this research.

REFERENCES

[1] An, S. H., Yeo, S. H., and Kang, Minsoo, “A Study on A Car Insurance Purchase Prediction Using Two-Class Logistic and Two-Class Boosted Decision Tree,” *Korean Journal of Artificial Intelligence,* 9-1, pp 9-14, 2021, DOI: http://dx.doi.org/10.24225/kjai.2021.9.1.9.

[2] Definisi / Penggunaan Kriteria Informasi Akaike (AIC) dalam Ekonometrika (2019, April 10). https://www.greelane.com/id/sains-teknologi-matematika/ilmu-sosial/introduction-to-akaikes-information-criterion-1145956.

[3] Dragos, C. M., Dragos, S. L. “Estimating Customers’ Behaviour in Motor Insurance Using Discrete Choice Model.” E a M : Ekonomie a management 20 (4) : 88-102, 2017, *DOI: 10.15240/tul/001/2017-4-007*

[4] Duan, Zhengmin, Chang, Y., Wang, Q., Chen, T., and Zhao, Qing, “A Logistic Regression Based Auto Insurance rate Making Model Designed for Insurance Rate Reform”. *International Journal of Financial Studies*, 2018, DOI:10.3390/ijfs6010018

[5] Guillen, M., Parner, J., Densgsoe, C., M, Ana., and Perez-Marin. Customer Loyalty in Insurance Industry : A Logistics Regression Approach. II Conference in Actuarial Science and Finance on Samos, 2002.