Jurnal Sains Matematika dan Statistika
p-ISSN: 2460-4542
e-ISSN: 2615-8663

PREDICTION OF PRODUCT PURCHASES THROUGH E-MAIL MARKETING CAMPAIGNS WITH LOGISTIC REGRESSION ANALYSIS

Ahmad Fahim Nashikhul Umam (11 pt Bold **)**

Data Science Study Program, Surabaya State University
Jl. Ketintang Sel. Kel, Ketintang, Kec. Gayungan, Surabaya, East Java 60231
Email: ahmadfahim.22022@mhs.unesa.ac.id
Author correspondence : ahmadfahim.22022@mhs.unesa.ac.id

Abstract

The study focuses on integrating logistic regression analysis in email marketing campaigns to predict consumer buying behavior. The data was gathered from a Kaggle dataset that included email opens, email clicks, discount offers, and purchase decisions. The data processing process involves conversion to Excel format, loading into R Studio, and identification of data types and missing data.

Logistic regression analysis was performed to identify factors influencing purchasing decisions, with model comparisons using criteria such as LR, log-likelihood, and AIC. The results suggest that the model may explain the variability of purchasing decisions, although not all predictor variables are significant. The prediction process with the best model is evaluated by comparing the prediction results with actual values.

The study provides an empirical foundation for the development of more rigorous email marketing strategies, illustrating the impact of variables such as age, email opens, email clicks, and discount offers. Although gender is not significant, other variables have an important role to play in understanding consumer buying behavior. This analysis can help marketing practitioners optimize their strategies to improve consumer response in the context of digital marketing.

Keywords: Marketing Campaigns, Digital Marketing, Email marketing, Logistic Regression.

1. Introduction

The rapid growth of information technology and internet penetration has changed the marketing landscape, presenting new opportunities to reach consumers. One method that is increasingly popular and effective is email marketing campaigns. Digital Marketing is a marketing activity carried out by brands to reach potential audiences on the internet using digital channels such as websites, emails, social media, and so on. Digital marketing is very different from conventional marketing, although as revealed above that the goal is the same. Especially in this article, we will discuss digital marketing through E-mail channels. This method can be used to provide the latest information about ongoing promotions or the latest products or services. In this context, prediction of purchasing behavior becomes a key aspect in designing a successful marketing strategy. This study aims to integrate logistic regression analysis in email marketing campaigns to predict product purchases by consumers. By leveraging historical purchase data and user information collected through email marketing campaigns [1] [2] . Digital marketing provides the ability to track and measure campaign results more accurately than traditional marketing methods [3] .

Logistic regression is a regression that uses two different values to express its response variable Usually used the value 0 to express failure and the value 1 to express success. Logistic regression analysis will be applied to identify critical factors influencing purchasing decisions. This is expected to provide deep insight into the characteristics of consumers who tend to make purchases after receiving marketing campaigns via email. [4]

The use of logistic regression as an analytical tool allows us to model the relationship between independent variables, such as product type, price, and campaign success, with the dependent variable, i.e. purchase decision. Thus, we can develop predictive models that can be used to amplify the effectiveness of email marketing campaigns, improve targeting, and ultimately increase sales conversions.

This research is expected to make a significant contribution in advancing our understanding of the dynamics of consumer behavior in the context of email marketing campaigns, as well as presenting an empirical foundation for the development of more careful and effective marketing strategies.

2. Research Methods

2.1 Data Collection

Data collection is a critical initial stage in this study. The dataset, taken from Kaggle's website, includes a variety of important information related to marketing campaigns, including data on whether emails are opened ('Emails Opened'), whether links in emails are clicked ('Emails Clicked'), whether there are discount offers ('Discounts Offered'), and consumer purchase decisions ('Purchased'). This data comes from a number of customers who have interacted with a marketing campaign via email.

2.2 Data processing

The initial step in data processing involves converting datasets into Excel format, enabling easy accessibility and data management using R Studio. After the conversion process, the dataset is loaded into the R Studio environment using the 'readxl' library, and the first few rows of data are displayed for initial inspection. The next step involves checking the data type to ensure the consistency of the variables. At this stage, identification of missing data is also carried out to understand the extent of the completeness of the dataset.

Next, a new data frame was created that focused on variables relevant for analysis, such as 'Purchased', 'Age', 'Gender', 'Email Open', 'Email Clicked', and 'Discount Offered'. This new data frame facilitates variable clarity and eases subsequent analysis steps. Additional data processing, such as handling outliers or normalizing variables, can also be performed as needed.

In the preliminary data analysis stage, descriptive statistics are carried out to provide a general overview of the distribution of these variables. Data visualizations, such as boxplots for consumer age distribution, can provide additional insights. The results of this stage of data processing provide a solid basis for subsequent analysis steps, including the application of logistic regression to predict consumer purchasing behavior. A mature understanding of dataset characteristics will support better interpretation of analysis results, aiding more informed decision making.

2.2 Logistic Regression

The next step of analysis in this study is to apply logistic regression to understand and model the relationship between response variables, i.e. purchasing decisions, with various predictor variables. Predictor variables considered include consumer age, gender,

whether the email was opened, whether links in the email were clicked, and whether there was a discount offer.

This application of logistic regression aims to identify the extent to which each predictor variable contributes to the likelihood of a purchase decision. Through this model, we can understand the relative impact of each predictor variable and the extent to which it can provide insight in designing more effective marketing strategies.

This process involves comparing the various possible logistic regression models, with different models including or ruling out certain predictor variables. Criteria such as the Akaike Information Criterion (AIC) can be used to evaluate and compare the relative advantages of each model. Models with lower AIC values are considered better, as they show good adjustments to the data without being too complex.

In addition to AIC, other statistical tests such as the likelihood ratio test can be used to evaluate the overall significance of the model. The results of this logistic regression analysis will provide an in-depth understanding of what factors are most influential on purchasing decisions, and thus, can be used to devise more targeted and efficient marketing strategies.

2.4 Predictions

After determining the best model of logistic regression based on evaluation criteria such as AIC, the next step is to apply the model to predict purchasing behavior on test data. Using the predictor variables that have been identified, the logistic regression model will generate the probability or likelihood of purchasing decisions for each observation on the test data.

This prediction process is very important because it provides an idea of the extent to which the model that has been developed is able to generalize the patterns that have been found in the training data into new data that has never been seen before. In other words, these predictions give an idea of the model's performance in real-world situations.

The predicted results will then be evaluated by comparing them with the actual values on the test data. This comparison can be made through various evaluation metrics such as accuracy, precision, recall, and F1-score, depending on the nature and specific purpose of the purchase prediction in the context of an email marketing campaign.

In addition, prediction error analysis can also provide valuable insights. Knowing where models tend to make mistakes can help refine the model or adjust marketing strategies more effectively. For example, does the model tend to miss potential customers or is there a situation where the model gives positive false signals.

Ultimately, this predictive step is a real implementation of the value that logistic regression models can generate. The results of a careful evaluation will make a significant contribution in assessing the reliability and relevance of the model in the practical context of email marketing, and can guide improvement and optimization steps going forward.

3. Results and Discussion

3.1 Descriptive Statistics

The contingency table of the categorical data can be seen in Table 1.

#	Open Email	
	Open	**No Open**
Purchased	5	5
No Purchased	5	5

Tabel 1. Contingency table between purchase and Openemail category predictor variables

#	Click Email	
	Click	**No Click**
Purchased	4	6
No Purchased	4	6

Tabel 2. Contingency table between purchase and Clickemail category predictor variables Tabel

#	Discount Offered	
	Discount	**No Discount**
Purchased	7	3
No Purchased	3	7

3. Contingency table between purchase and discount offered category predictor variables

From Table 1, it can be seen that opening an email or not does not seem to influence the purchase decision.

From Table 2, clicking on an email may have an impact on purchasing decisions, as the proportion of purchases differs between those who click and don't click.

From Table 3, discount offers seem to have a positive impact on purchasing decisions, with more making purchases when discounts are offered.

While the summary of predictor variables with ratio data is contained in Table 4.

Variable	Min	Q1	Median	Mean	Q3	Max
Age	15.00	24.50	38.00	38.25	53.50	62.00
Gender	0.00	0.00	1.00	0.55	1.00	1.00

Table 4. Summary of predictor variables with ratio data

The correlation of all variables is shown in figure 1 as follows:

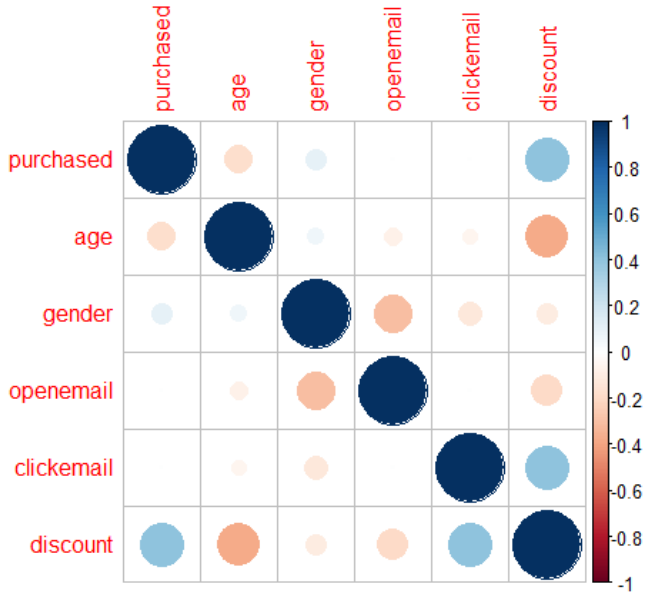

Figure 1, plots the correlation matrix of all variables.

From Figure 1, it can be seen that the correlation coefficient between variables is Variations. The correlation between purchases and gender is quite low, whereas with discount offers has a positive correlation of about 0.8.

3.2 Modelers with logistic regression

In this regression, the link function will use logits. The data will be analyzed with several possible models with the following 4 Tables.

Coefficients	Estimate	Std. error	z-value	p-value
Intercept	-3.061017	3.779512	-0.810	0.418
Age	0.007953	0.060742	0.131	0.896
Gender	1.914984	1.470734	1.302	0.193
Open Email	0.442313	1.396918	0.317	0.752
Click Email	0.606280	1.759349	0.345	0.730
Discount	2.056277	2.242488	0.917	0.359

Table 4. Parameter estimation using logistic regression.

Statistical probability ratio (LR) = 4.5002 (P-value = 0.4799), log-likelihood = -8.8402 (for Model 1) and -11.0904 (for Model 2), and AIC value = 29.68. The P value of the ratio

statistic is likely greater than the significant level of 0.05, indicating that there is not enough evidence to reject the null hypothesis, which means there is no significant difference between Model 1 and Model 2. The values in Table 4 indicate that the predictor variables (age, gender, openemail, clickemail, discount) did not contribute significantly to explaining variability in the response variable (purchased). Therefore, based on the likelihood ratio test analysis, a model containing only constants (Model 2) is not significantly worse at explaining data compared to the more complex Model 1.

In other words, based on the results of these statistical tests, there is not enough evidence to state that the independent variable has a significant impact on the response variable. Therefore, we cannot conclude that the independent variable significantly affects the likelihood of purchase based on the results of this likelihood ratio test.

A summary of the phased process is listed in Table 5.

Step	Predictor	AIC
1	Age, Gender, Open Email, Click Email, Discount	29.68046
2	Age, Gender, Open Email, Click Email	28.62987
3	Age, Gender, Open Email	28.4141
4	Age, gender	26.41971
5	Age	25.78498

Table 5. Parameters of some models using logistic regression

From Table 5 it can be seen that model 5 has the smallest AIC among other models. So Table 6 will provide parameter estimates for model 5.

Coefficients	Estimate	Std. Error	Z-value	P-value
Intercepts	--1.854577	2.070015	-0.896	0.370
Age	0.004907	0.033987	0.144	0.885
Gender	1.133815	0.836895	1.355	0.175
OpenEmail	0.299478	0.844822	0.354	0.723
ClickEmail	0.297273	1.032609	0.288	0.773
Discount	1.279272	1.279405	1.000	0.317

Tabel 6. Estimated Parameters for the best model using logistic regression

Logistic regression analysis shows statistically relevant results for evaluating the model. The statistical probability ratio (LR) reaches a value of 13.24, with a p-value of 0.0211. This value indicates that there is at least one significant predictor variable in the model. These results provide an indication that this logistic regression model may make a meaningful contribution in explaining purchasing decision variability.

Furthermore, the log-likelihood obtained from the model is -43.283, and the AIC (Akaike Information Criterion) value is 98.566. This AIC value can be used as a criterion to compare the model with other models, where a lower AIC value is considered better. With the P value of statistically significant probability ratios, the relatively low AIC value suggests that the model may be quite good at explaining consumer purchasing data.

Looking at the Coefficients table, predictor variables, such as age, gender, open-opens, click-in links, and discounts, all have different estimated coefficients. However, not all predictor variables have a significant influence on purchasing decisions. For example, gender has a p-value of 0.175, suggesting that gender may not play a significant role in predicting purchases.

However, these results do not guarantee that this model is the best model, because there may be multicollinearity among the predictors. Therefore, further analysis is carried out by applying gradual regression to ensure the reliability and validity of this model. Such processes provide further understanding of the relationships of predictor variables and can refine models to improve predictions of consumer buying behavior.

3. Conclusion

In conclusion, this study explores the integration of logistic regression analysis in email marketing campaigns to predict consumer buying behavior. Data was collected from Kaggle's dataset, covering variables such as email opens, email clicks, discount offers, and purchase decisions. The data processing process involves converting to Excel format and loading into R Studio, focusing on checking data types and identifying missing data.

Logistic regression analysis is performed to identify factors influencing purchasing decisions, comparing models with criteria such as LR, log-likelihood, and AIC. The results suggest that the model may explain the variability of purchasing decisions, although not all predictor variables are significant. The prediction process uses the best model, and evaluation is done by comparing the prediction results with actual values.

This study provides an empirical foundation for the development of more careful email marketing strategies. While gender is not significant, other variables such as age, email opens, email clicks, and discount offers have a different impact. This analysis can help marketing practitioners design strategies that are more targeted and responsive to consumer preferences in the context of digital marketing.

Bibliography

[1] A. Esti, P. Faculty, S. Vocational, U. Science, T. Computer, and F. Lupiana, "The Role of Marketing Information Systems in Managing Marketing Processes Through Digital Marketing," *Economics and Business*, vol. 2, no. 2, pp. 88–102, 2023, [Online]. Available: http://ejurnal.provisi.ac.id/index.php/JIMEB

[2] E. Mulyantomo, A. I. Sulistyawati, and D. Triyani, "Online Marketing and Digital Branding Training During the Covid-19 Pandemic for MSME Actors in Tegalarum Village, Mranggen District, Demak Regency," 2021. [Online]. Available: https://journals.usm.ac.id/index.php.tematik

[3] L. Rahmawati, M. Ikaningtyas, J. Rungkut Madya No, and J. Timur, "Application of Digital Marketing to Support MSMEs Sebitt Snack in Kebumen Application of Digital Marketing to Support SME Bites Snack in Kebumen," *JIPM:Journal of Community Service Information*, vol. 1, no. 3, pp. 63–71, 2023, doi: 10.47861/jipm-nalanda.v1i3.310.

[4] S. Mayadi, "UNIVARIATE LOGISTIC REGRESSION WITH UNBALANCED RESPONSE DATA," 2009.