Dean&Francis

ISSN 2959-6157

Analyzing the Impact of Customer Engagement on Conversion in Digital Marketing Scenarios based on Multiple Linear Regression and Random Forest Model

Yanlin Lu[1, *]

[1]Shanghai Datong High School, Shanghai, 200000, China

*Corresponding author: 200237@yzpc.edu.cn

Abstract:

This study explores the impact of customer engagement on conversion rates in the context of digital marketing scenarios. Different customer engagement indicators are studied in depth by integrating multiple linear regression and random forest model. A dataset containing demographic information, marketing variables, and customer engagement variables was obtained from Kaggle, and the model was constructed with customer engagement variables as explanatory variables and conversion as response variable. The results of the multiple linear regression model showed that variables other than social sharing had a significant positive effect on transformation, passed the F-test, and had no covariance or auto-correlation problems, but data normality was not fully satisfied. The random forest model was accurate and fitted well on the test set. The study shows that there are differences in the order of importance of customer engagement indicators in different models, and more accurate conclusions need to be analyzed in combination with practical application scenarios, so as to provide guidance for marketers to understand the relationship between customer engagement and conversion and formulate relevant strategies.

Keywords: Multiple linear regression; random forest; customer engagement; conversion; digital marketing.

1. Introduction

In today's era of globalization, the rapid development of digital technology has profoundly changed the world's business landscape. The global economy is constantly transforming and upgrading under the wave of digitization, and various industries are actively exploring how to utilize digital channels to

expand their business and enhance their competitiveness.

Thakur and Das proposed the application of SPSS software in market research, including conducting accurate data analysis, identifying market trends, predicting consumer behavior, and developing effective marketing strategies [1]. There are various approaches to data analysis in business that can be selected based on data and specific research topics. For example, Shi et al. proposed an improved linear regression algorithm including ridge regression, Lasso regression, and elastic network regression, introduced the process of business behavior analysis including data collection, data preprocessing, data analysis, and data visualization, and validated the algorithm by applying it in the analysis of the sales data of an e-commerce company [2].

With the development of the Internet and the maturity of social media and online shopping software, digital marketing occupies an increasing proportion of marketing. As a key area in this change, digital marketing has gained widespread attention and application worldwide.

In the field of digital marketing, numerous companies and marketers are working to improve the efficiency and effectiveness of their marketing campaigns. A deterministic factor analysis was employed, specifically using chain of alternatives, exponentials and integrals to reveal the isolated effects of factors as well as determining the order of the factors in the cofactor chain [3]. Factor analysis is used by Roy et al. to identify factors that influence consumer behavior. Specifically, Kaiser-Meyer-Olkin (KMO) and Bartlett tests were included to verify the suitability of the variables for factor analysis, followed by principal component analysis with Varimax Orthogonal Rotation to determine the optimal factor structure [4]. In the general environment of expanding research in the field of digital marketing, there have been related studies that provide valuable references and insights for this paper.

For example, Zhang et al. constructed a model of the impact of customer engagement on stickiness based on the theories of customer engagement, value co-creation and relationship marketing, and analyzed it using structural equation modeling to conclude that customer engagement has a direct positive impact on customer stickiness as well as an indirect impact through customer value creation [5]. This reveals that the role of customer engagement in digital marketing scenarios cannot be ignored and its impact may be multifaceted, which provides a similar research idea to explore the impact of customer engagement on conversion in digital marketing, i.e., customer engagement may act on the core metrics that this paper focuses on through multiple pathways. Hampton et al. focused on customer engagement in the retail electricity market. The study not only establishes a definition of customer engagement in retail electricity markets, but also provides a systematic overview of these customer engagement strategies in retail electricity markets. It reminds people of the unique research value of customer engagement in different domains and the need to clarify the related concepts and key influences [6].

In summary, these studies provide guidance and implications for analyzing the impact of customer engagement on conversion rate in digital marketing scenarios based on multiple linear regression in a variety of ways, ranging from research methodology, research ideas to key factor analysis.

Conversion rate is one of the key indicators to measure the success of digital marketing campaigns. A high conversion rate means that an organization can convert more potential customers into actual purchasers or long-term loyal users, thus achieving business growth and profitability. With the rise of digital platforms such as social media and online communities, the forms of customer engagement have become increasingly diverse. Customers are no longer just recipients of information, but are able to actively participate in the marketing process, influencing marketing results by posting comments, sharing content, and engaging in interactions. However, the relationship between customer engagement and conversion has not yet been fully and deeply understood.

The topic of this paper is the effect of customer engagement on conversion, which belongs to multiple independent variables working together to influence the dependent variable. Multiple linear regression model is widely used in this logistic relationship, with more established examples of practice in several fields. Popescu et al. used multiple linear regression algorithms to construct easily interpretable and understandable models to explore the impact of the type of activity students engage in on social media tools on the final learning outcomes [7]. Chen et al. examined the factors influencing the price of housing in Beijing to help people assess home purchase expectations, considered the interaction effects between variables, and solved the covariance problem by adding an interaction term and using Forward Stepwise Regression to make the model more accurately reflect the influencing factors of housing prices [8]. Ju et al. helped college students to understand the impact of different test scores on their chances of applying for graduate school admissions, so as to determine test preparation priorities, using a combination of simple linear regression, multiple linear regressions, and stepwise regression, the data were analyzed in depth, and the model was validated and optimized through t-tests and regression diagnostics [9].

Focusing on digital marketing scenarios, this study adopts multiple linear regression and random forest, aiming to

ISSN 2959-6157

deeply explore the impact of customer engagement on conversion. Through the collection and analysis of relevant data, it tries to establish a quantitative relationship model between customer engagement factors and conversion rate, so as to provide valuable theoretical basis and practical guidance for enterprises on how to effectively improve customer engagement and thus conversion rate in digital marketing activities.

2. Methods

2.1 Data Source

The dataset from Kaggle is used to predict conversion in digital marketing. The data consists of demographic information, marketing-specific variables, customer engagement variables, historical data, and target variable. Potential applications are predictive modeling of customer conversion rates, analyzing the effectiveness of different marketing channels and campaign types, identifying key factors driving customer engagement and conversion, optimizing ad spend and campaign strategies to improve Return On Investment (ROI) [10].

2.2 Variable Selection

The research topic of this paper is identifying key customer engagement factors driving conversion, so this paper can choose customer engagement variables as the explanatory variables and conversion as the response variable. The meanings of the variables are as follows (Table 1):

Table 1. Variable introduction

Variable	Logogram	Meaning
Website Visits	x_1	Number of visits to the website
Pages Per Visit	x_2	Average number of pages visited per session.
Time On Site	x_3	Average time spent on the website per visit (in minutes)
Social Shares	x_4	Number of times the marketing content was shared on social media
Email Opens	x_5	Number of times marketing emails were opened
Email Clicks	x_6	Number of times links in marketing emails were clicked
Click Through Rate	x_7	Rate at which customers click on the marketing content
Conversion Rate	x_8	Rate at which clicks convert to desired actions (e.g., purchases)
Previous Purchases	x_9	Number of previous purchases made by the customer
Loyalty Points	x_{10}	Number of loyalty points accumulated by the customer
Conversion	Y	Binary variable indicating whether the customer converted (1) or not (0)

2.3 Method Introduction

2.3.1 Multiple linear regression

In this paper, the multiple linear regression model is used first, which has the advantage of having a clear mathematical expression, and the coefficients can intuitively represent the direction and degree of the influence of each independent variable on the dependent variable; it has a certain degree of tolerance for the slight deviation of the data from these assumptions; and the computational process is relatively simple and mainly involves matrix arithmetic, with a low degree of complexity, so that it can quickly complete the training and prediction of the model. However, it has certain limitations. Multiple linear regression assumes that there is a linear relationship between the dependent and independent variables and that the error term satisfies the conditions of normality, independence and homoscedasticity. In practical problems, these assumptions may not be fully valid. If there are nonlinear relationships, interactions, or outliers in the data, multiple linear regression may not fit the data well, resulting in lower values. When analyzing the relationship between customer behavior and purchase intention, the customer's purchase decision may be influenced by the interaction of multiple complex factors, and the relationship between these factors may not be a simple linear relationship.

2.3.2 Random forest

Random forest is an integrated learning method based on decision trees, which can automatically capture the

nonlinear relationships and interactions between independent variables. Compared with multiple linear regression, random forests do not require strict assumptions about the distribution of the data and the relationships between the variables. By constructing multiple decision trees and integrating them, random forests can often achieve higher predictive accuracy than a single model. In many cases, when linear models perform poorly, random forests can improve the performance of the model by learning complex patterns in the data.

3. Results and Discussion

3.1 Multiple linear regression

The analysis in this paper shows that there are many factors influencing conversion. As Figure 1 shows:

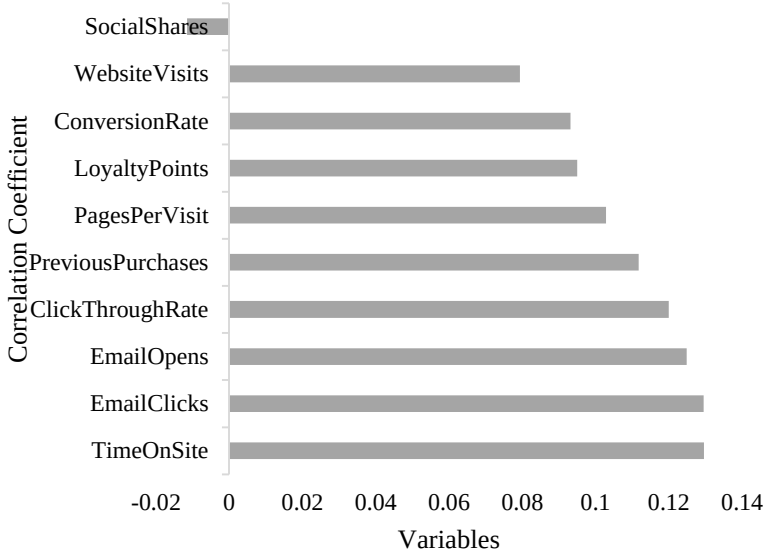

Fig. 1 Pearson correlation analysis

At this point, a rough correlation between the data can be identified. The positive and negative values of Pearson's coefficient can clearly reveal whether the variables are positively or negatively linearly correlated, and the absolute value of the coefficient reflects the stronger the linear correlation between the variables. For multiple linear regression, the Pearson coefficient test is helpful in many ways. It assists in detecting multicollinearity problems. Through the Pearson's coefficient test, if the absolute value of the coefficients between some independent variables is close to 1, then this paper needs to be alert to the fact that these independent variables may cause multicollinearity problems, so as to deal with the variables appropriately before the regression analysis. Secondly, it helps in the initial assessment of the importance of the variables. Those independent variables with high absolute values of Pearson coefficients with the dependent variable may have a greater impact on the dependent variable in multiple linear regression, which can provide a basis for prioritizing these variables in constructing the regression model.

In addition, the results of the Pearson coefficient test and the multiple linear regression results are compared with each other, and if they are consistent, the relationship between the variables can be further confirmed; if they are not consistent, it suggests that there may be other influencing factors, which prompts people to further explore the underlying mechanisms behind the data.

After analyzing the Pearson correlation matrix for each factor, this paper can perform multiple regression analysis. The formula is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{10} x_{10} + \epsilon \quad (1)$$

The linear regression model has certain conditions for use: linear relationship, normality, homoscedasticity, residual independence, and no covariance. The test methods and determination results of these conditions in this paper will be explained individually later. The results of the multiple linear regression are shown in the table 2 below:

ISSN 2959-6157

Table 2. Multiple Linear Regression Analysis

	Unstandardized coefficient		Standardized coefficient	t	p	Covariance diagnostics	
	B	S.E.	Beta			VIF	Tolerance
Constant	0.296	0.02	-	14.802	0.000**	-	-
Website Visits	0.002	0	0.085	8.019	0.000**	1.002	0.998
Pages Per Visit	0.013	0.001	0.103	9.774	0.000**	1.001	0.999
Time On Site	0.01	0.001	0.132	12.543	0.000**	1.001	0.999
Social Shares	0	0	-0.009	-0.823	0.41	1.001	0.999
Email Opens	0.007	0.001	0.125	11.862	0.000**	1	1
Email Clicks	0.015	0.001	0.13	12.31	0.000**	1	1
Click Through Rate	0.5	0.041	0.128	12.104	0.000**	1.001	0.999
Conversion Rate	0.559	0.063	0.093	8.845	0.000**	1.001	0.999
Previous Purchases	0.013	0.001	0.114	10.789	0.000**	1.001	0.999
Loyalty Points	0	0	0.099	9.4	0.000**	1.001	0.999
R-square	0.113						
F	F (10,7989)=101.866,p=0.000						
D-W value	1.72						
* p<0.05 ** p<0.01							

Table 2 shows that the model equation is:

$$y = 0.296 + 0.002x_1 + 0.013x_2 + 0.01x_3 - 0.00x_4 + 0.007x_5 + 0.015x_6 + 0.5x_7 + 0.559x_8 + 0.013x_9 + 0.000x_{10} + \epsilon \quad (2)$$

The final analysis shows that Social Shares does not affect Conversion and the rest of the variables have a significant positive effect on Conversion. The model R-squared value is 0.113, which means that the independent variable can explain the dependent variable's 11.3% of the cause of change. The F-test of the model found that the model passes the F-test (F=101.866, p=0.000<0.05), which means that at least one of the independent variables will have an impact on Conversion.

In addition, the multiple covariance test of the model reveals that the VIF values of the model are less than 5, indicating that there is no problem of covariance, and the D-W is around 2, indicating that there is no auto-correlation in the model, and there is no correlation between the sample data, so the model is relatively good.

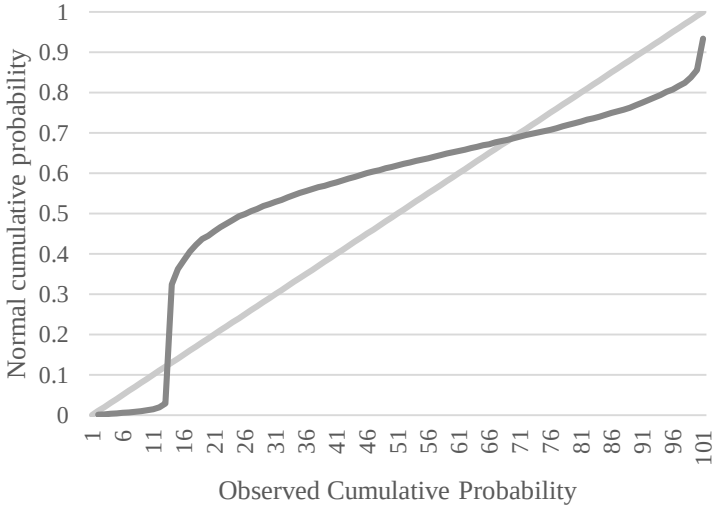

Fig. 2 Normalized P-P plots of regression standardized residuals

P-P plots are often used to visualize whether data are normally distributed (Figure 2). The principle is that if the data is normal, then the cumulative proportion of the data is basically the same as the cumulative proportion of the normal distribution. Use the actual data accumulation ratio as the X-axis and the corresponding normal distribution accumulation ratio as the Y-axis to make a scatter plot. If the scatter plot appears as a diagonal line, then the data are normally distributed. Conversely, the data is not normal. From Figure 2, it can be found that the P-P plot of the residuals shows a certain trend but more points in the data do not lie on a straight line, thus indicating that the data do not quite satisfy the quality of normality. However, the regression analysis in this paper only establishes the relationship between X and Y, without the need to predict the Y value credibility, the criteria of normality can be appropriately loosened.

The scatterplot of residual diagnosis will take its predicted value as X-axis and the residual value as Y-axis, if all the points are uniformly distributed on both sides of the straight line Y=0, then it can be considered to satisfy variance chi-square. Figure 3 shows that there is a clear clustering of three groups of points and the trend lines formed by these three groups are parallel, with distinct grouping characteristics, indicating that there may be some natural grouping of the data. This may be due to different customer groups, different marketing channels and types causing this grouping.

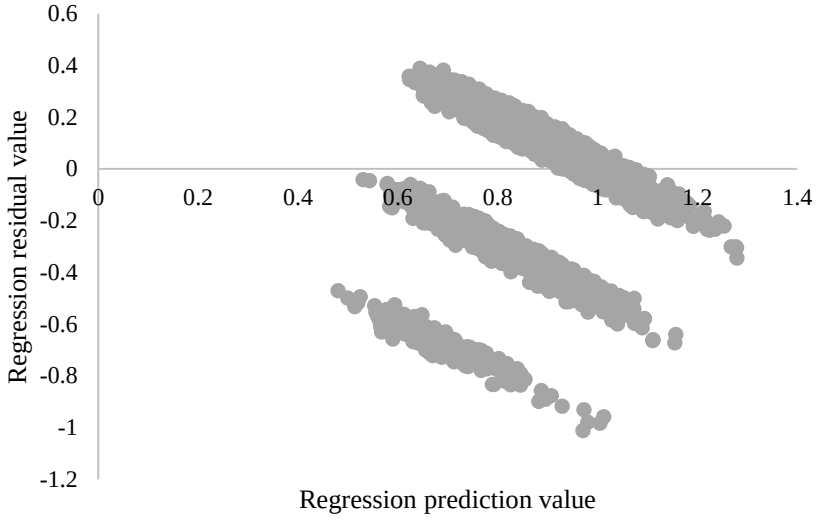

Fig. 3 Residual Diagnostic Graph

Table 3. Model Evaluation

R	R square	Adjusted R square	Model error RMSE	DW	AIC	BIC
0.336	0.113	0.112	0.31	1.72	3978.278	4055.137

According to Table 3, taken together, this model has some advantages, such as high overall significance and weak autocorrelation of residuals, but it also has some shortcomings, such as weak explanatory ability of the dependent variable and the prediction accuracy needs to be improved, but this does not have an impact on the research theme of this paper.

3.2 Random forest

Random forest model was performed with the 10 variables in Table 1 as the independent variables and transformation as the dependent variable, with the training set scale set at 0.8, the number of decision trees at 100, and the maximum depth of the trees unrestricted. The final model obtained 89.88% accuracy, 88.72% precision (combined), 89.88% recall (combined) and 0.87 f1-score (combined) on the test set. The model results are acceptable.

ISSN 2959-6157

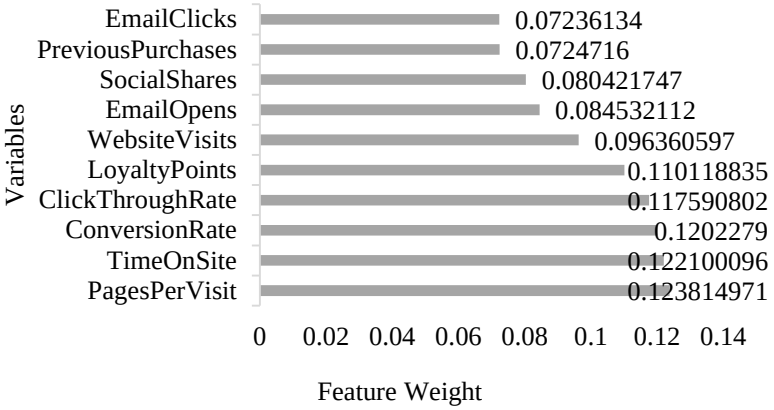

Fig. 4 Feature weight graph

Figure 4 shows the importance of the contribution of each heading to the model.

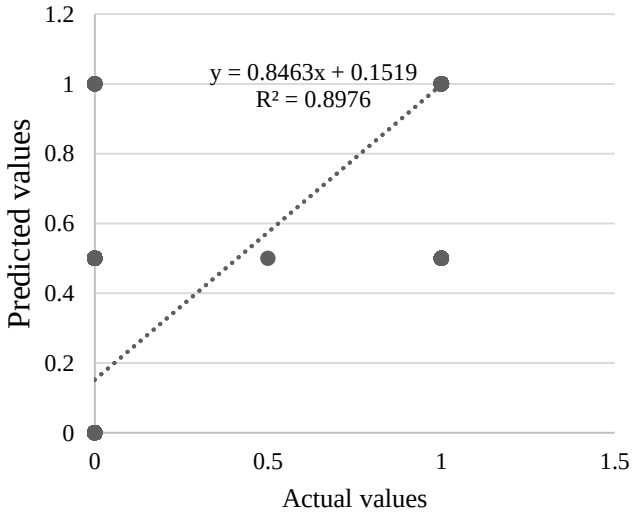

Fig. 5 Scatter plot of the actual values against the predicted values

Figure 5 shows that the final R-square value obtained is 0.90, indicating that the model fits the data well and has a strong explanatory ability for the dependent variable.

3.3 Discussion

The aim of this study was to analyze the impact of customer engagement on conversion rates in Digital Marketing Scenarios by using multiple linear regression and random forest models. Through these models, the author was able to rank the importance of several indicators of customer engagement. However, it is worth noting that the two models do not yield the same results, so the specific conclusions need to be analyzed in the context of actual application scenarios.

The result of Figure 4 and Figure 6 can help marketers prioritize their efforts and allocate resources more effectively. Among these indicators, website visits and click-through rates are relatively important factors. Marketers can use the results to refine their marketing strategies, such as optimizing ad texts, improving targeting, and enhancing the visual appeal of marketing campaigns, thus bringing more traffic to digital platforms and increasing conversion opportunities.

Social media sharing presents its relatively insignificant impact on conversion in both graphs. It may be more of a reflection of a user's desire to express themselves, their need for social interaction, or their approval of the content, but it doesn't directly equate to purchase intent or conversion behavior. Users may share content because they find it interesting, valuable, or to showcase their interests, but they themselves may not necessarily make a purchase in the immediate or short-term.

Fig. 6 Multiple linear regression standardized coefficients

4. Conclusion

In conclusion, although this study provides some insights into the importance of certain customer engagement metrics, it is undeniable that the relationship between customer engagement and conversion exhibits a high degree of complexity, e.g., different customer segments as well as different marketing channels may have a significant impact on the relationship.

In the actual digital marketing scenarios, choosing the right model is crucial in order to effectively analyze and predict customer behavior. However, there is no universal model that can be perfectly applied to all situations, and its applicability often depends on specific practical application scenarios. In a scenario with relatively simple variable relationships, a more basic model can meet the needs, while in the case of large-scale data, many variables and complex relationships, more complex and advanced models are required.

Future research could focus on exploring different models and approaches to obtain more accurate predictions, thus providing digital marketers with more insights of practical operational value to help them make more informed decisions when developing, executing and optimizing their marketing strategies.

References

[1] Thakur J, Das P. Deploying SPSS in Digital Market Analytics: Leveraging it tools for informed Business Strategies. 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS), 2023, 1277-1282.

[2] Shi Y. Application of improved linear regression algorithm in business behavior analysis. Procedia Computer Science, 2023, 228: 1101-1109.

[3] Semenov V P, et al. Factor analysis of the results of digital technology applications in the company's marketing activities. 2017 XX IEEE International Conference on Soft Computing and Measurements (SCM), 2017, 879-882.

[4] Roy B, Acharjee P B, Ghai S, Ghai S, Shukla A, Sharma N. Impact of digital media marketing on consumer buying decisions. 2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies, 2024, 26: 1-5.

[5] Zhang M, Guo L, Hu M, Liu W. Influence of customer engagement with company social networks on stickiness: Mediating effect of customer value creation. International Journal of Information Management, 2022, 37(3): 229-240.

[6] Hampton H, Foley A, Del Rio D F, Smyth B, Laverty D, Caulfield B. Customer engagement strategies in retail electricity markets: A comprehensive and comparative review. Energy Research & Social Science, 2022, 90: 102611.

[7] Popescu P S, Mihaescu M C, Popescu E, Mocanu M. Using ranking and multiple linear regression to explore the impact of social media engagement on student performance. 2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT), 2016, 250-254.

[8] Chen B, Min Y, Yu S. The research on factors influencing housing prices-take Beijing as an example. Highlights in Science, Engineering and Technology, 2024, 88: 724-730.

[9] Ju W, Wang H, Ye Z. Analysis of relative importance of factors affecting the chance of admission of applying to graduate students. Journal of Education, Humanities and Social Sciences, 2022, 6: 99-109.

[10] Leng Jianfei, Gao Xu, Zhu Jiaping. Application of multiple linear regression statistical prediction model. Statistics and Decision Making, 2016, 7: 82-85