ESTUDIOS DE ADMINISTRACIÓN
ISSN 0719-0816 versión en línea

Predictive Modeling of Customer Purchase Behavior in Social Media Advertising: A Logistic Regression Approach

Keywords: Customer, Social Media, Logistic Regression Approach.
Received: 2023 | **Accepted:** 2024 | **Available online:** 2025
Cite this article as: Barrera Legorburo, L. (2024). *Predictive Modeling of Customer Purchase Behavior in Social Media Advertising: A Logistic Regression Approach. Estudios de Administración*, 31(1), 39–48.
https://doi.org/10.5354/0719-0816.2024.79990

Lisgrey Barrera Legorburo
Universidad de Chile
lbarreral@fen.uchile.cl

ABSTRACT

This study investigates the impact of demographic factors on purchase behavior in social media advertising, addressing a key issue for marketers: identifying which characteristics can enhance targeted ad strategies. Using logistic regression, the research examines how age and estimated salary influence the probability of making a purchase, offering insights into consumer decision-making in digital environments. The analysis draws on a dataset of 400 observations from a survey of active social media users across platforms like Instagram, Facebook, and Twitter. A logistic regression model was trained to assess the relationship between demographic predictors and purchase outcomes, with subsequent testing for predictive accuracy. Both age and estimated salary emerged as significant predictors, with each showing a positive association with purchase probability. Marginal effects analysis highlighted the stronger influence of age on purchase likelihood, while estimated salary, though statistically significant, showed a subtler effect. Additionally, odds ratios confirmed the predictive strength of these factors. Model performance was evaluated using accuracy, precision, and recall metrics derived from a confusion matrix, demonstrating high reliability in predicting purchasers and non-purchasers, albeit with a conservative tendency. The distribution of predicted probabilities indicated strong confidence in classifying non- purchasers, supporting the model's cautious approach to positive predictions. These findings provide practical insights for marketers seeking to optimize ad targeting by leveraging demographic data. By understanding the demographic drivers of purchase decisions in social media contexts, this study contributes to the development of more efficient and effective advertising strategies, ultimately enhancing customer engagement. Future research could expand this model by incorporating additional demographic or psychographic variables, facilitating a more nuanced approach to predicting purchase behavior in digital advertising.

Keywords: Customer, Social Media, Logistic Regression Approach.

JUSTIFICATION

The rapid growth of social media as a marketing channel has fundamentally changed how companies connect with consumers, creating an urgent need for precise, data-driven strategies to understand and influence purchase behavior (Naem & Okafor, 2019). Demographic factors, especially age and income, are well-established as critical in shaping consumer decisions due to their links with purchasing power, preferences, and responsiveness to marketing stimuli (Kotler & Keller, 2016; Solomon, 2018). However, while traditional retail contexts have been extensively studied, there is a pressing need for research focusing on the unique dynamics of social media advertising, where ad exposure and consumer engagement differ substantially from other platforms (Kaplan & Haenlein, 2009; Akar & Topçu, 2011).

This study addresses this gap by using logistic regression to quantitatively evaluate the effects of age and estimated salary on purchase likelihood within social media environments. By examining these demographic predictors, the research enhances our understanding of consumer behavior in digital advertising and provides marketers with actionable insights to improve ad targeting. This focus is increasingly relevant as firms face greater demands to justify marketing expenditures and optimize return on investment (ROI) through data-driven campaigns (Chaffey & Ellis-Chadwick, 2019; Stieglitz et al., 2017).

RESEARCH OBJECTIVE

This study aims to investigate the impact of demographic factors—specifically age and estimated salary—on purchase probability within social media advertising. It tests two hypotheses: (1) Age and estimated salary significantly predict consumer purchasing behavior, and (2) Marginal effects and odds ratios clarify the influence of these factors on purchase likelihood. The findings are intended to enhance targeted advertising strategies based on demographic data in digital platforms.

METHODOLOGY

This study applies to a logistic regression model to assess the impact of age and estimated salary on the likelihood of consumer purchase within the context of social media advertising. Given the binary outcome variable (Purchased: 1 = Purchase, 0 = No Purchase), logistic regression is well-suited for estimating the probability of purchase, as it effectively links predictor variables to the log-odds of an outcome (Hosmer et al., 2013; Menard, 2010). This approach aligns with logistic modeling theory, allowing for a clear interpretation of predictors' effects through log-odds and odds ratios, enhancing understanding of the demographic factors influencing purchasing behavior (Peng et al., 2002) , as represented by the following formula:

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{EstimatedSalary}$$

The dataset consists of 400 observations from a 2023 survey targeting active social media users on platforms such as Instagram, Facebook, and Twitter. These observations include demographic and economic information for consumers exposed to social media ads. Data preprocessing involved calculating summary statistics, checking for missing values and duplicates, and visualizing the distribution of age and estimated salary by purchase status. To ensure effective model evaluation, the dataset was partitioned into training and test sets, with a random assignment of 90% of observations to the training set and 10% to the test set, yielding 365 observations (91.25%) for training and 35 (8.75%) for testing (Stieglitz et al., 2017). This partitioning strategy provides a robust foundation for model training and evaluation, as shown in Table 1.

Table 1. Data Partitioning for Training and Testing.

Train Set	Frequency	Percent	Cumulative Percent
0 (Test)	35	8.75	**8.75**
1 (Train)	365	91.25	**100.00**
Total	**400**	**100**	**100**

The logistic regression model was fitted on the training subset to estimate the effects of age and estimated salary on purchase probability, with marginal effects and odds ratios calculated to improve interpretability of results. Model performance was evaluated on the test subset using accuracy, precision, and recall metrics derived from the confusion matrix, offering a comprehensive view of the model's classification effectiveness. This rigorous approach, grounded in logistic modeling theory, combines theoretical and practical insights to support data-driven marketing strategies effectively (Hosmer et al., 2013; Peng et al., 2002).

RESULTS

This section presents the results in three parts: (1) descriptive statistics, (2) hypothesis testing and model estimates, and (3) performance metrics and interpretation of the model's predictive accuracy.

Descriptive Statistics

The descriptive statistics in Table 2 summarize the sample's characteristics in terms of age, estimated salary, and purchase behavior. The average age is 37.7 years with a standard deviation of 10.5, indicating a diverse age representation from 18 to 60. Estimated salary averages $69,742.50, with high variability (standard deviation of $34,096.96) and a range from $15,000 to $150,000, reflecting a broad spectrum of economic backgrounds. The binary purchase variable shows a mean of 0.3575, indicating that about 35.8% of individuals made a purchase. This baseline provides an initial insight into the likelihood of purchase within this dataset, which will be further analyzed to determine how age and estimated salary influence purchase behavior. Data quality checks indicate that there are no missing values or duplicate records, confirming the integrity of the dataset for further analysis. These initial descriptive statistics set the foundation for subsequent analyses to

explore the relationships between demographic factors and purchasing decisions.

Table 2. Descriptive Statistics.

Variable	Mean	Std. Deviation	Min	Max
Age	37.655	10.48288	18	**60**
Estimated Salary	69742.5	34096.96	15000	**150000**
Purchased (binary)	.3575	.479864	0	**1**
Missing Values	**None**			
Duplicate Records	**None**			

The data visualizations offer insights into the distributions of age and estimated salary by purchase status, revealing demographic trends that may influence purchase likelihood. In Figure 1, purchasers are generally older than non-purchasers, indicating a positive relationship between age and purchase probability. Similarly, Figure 2 shows that individuals with higher estimated salaries are more likely to make a purchase, supporting the hypothesis that income positively impacts purchase behavior. Figure 3 reveals that approximately 36% of the sample made a purchase, providing a baseline purchase probability. These visualizations reinforce the inclusion of age and estimated salary as predictors in the logistic regression analysis, suggesting both factors play a significant role in consumer purchase decisions in social media advertising.

Figure 1. Age Distribution by Purchase Status.

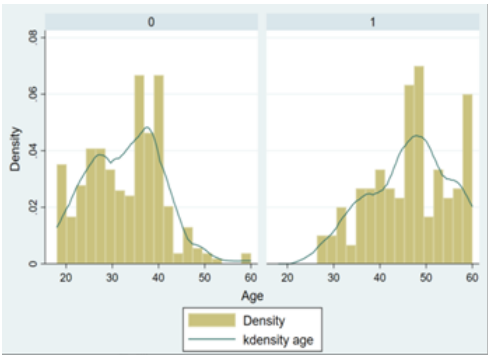


Figure 2. Estimated Salary Distribution by Purchase Status.

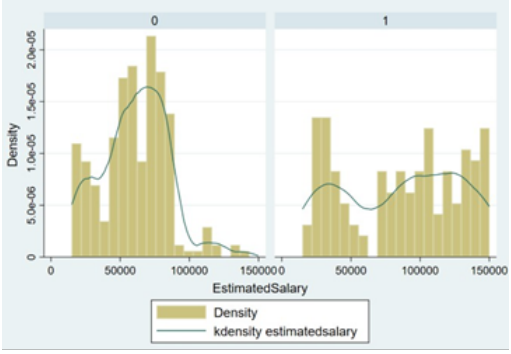
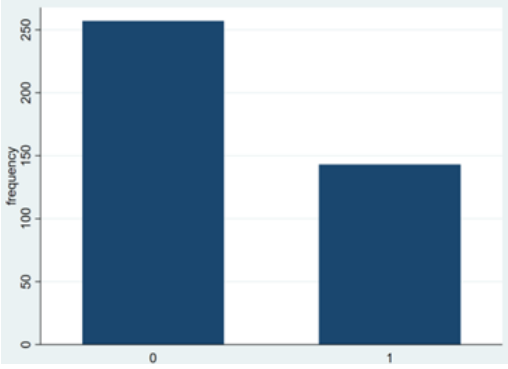


Figure 3. Overall Purchase Distribution.



Hypothesis Testing and Model Estimates

The logistic regression model estimated purchase probability based on Age and Estimated Salary, both of which showed statistically significant coefficients, affirming that these factors positively influence purchase likelihood. As indicated in Table 3, the Age coefficient (0.2239, $p < 0.001$) suggests that each additional year increases the log-odds of purchase, while the Estimated Salary coefficient (0.0003, $p < 0.001$) points to a positive association between higher salary and purchase probability. The model converged after five iterations with a final log likelihood of -131.3114. Model fit was assessed with a pseudo-R^2 of 0.4502 and an LR chi-square of 215.06 ($p < 0.001$), indicating it explains a substantial portion of purchase behavior variance. As noted in Table 3, an initial estimation was followed by up to 500 iterations to ensure stability and convergence.

Table 3. Logistic Regression Results.

Variable	Coefficient (β)	Std. Error	z-value	P>|z|	95% CI
Age	0.2239	0.0257	8.72	0.000	**[0.1736, 0.2742]**
Estimated Salary	0.0003	5.52e-06	6.32	0.000	**[0.0002, 0.0005]**
Intercept	-11.9784	1.2928	-9.27	0.000	**[-14.5124, -9.4445]**

Model Statistics
Log likelihood -131.3114
LR chi² (2) 215.06
Pseudo R² 0.4502
Observations 365
Note: The initial estimation was performed, and the model was subsequently iterated up to 500 times to ensure convergence.

Table 4 presents the odds ratios for the predictors, offering a more interpretable perspective on the model's estimates. The odds ratio for Age was 1.2509 (95% CI: [1.1895, 1.3155]), implying that each additional year of age increases the odds of purchase by approximately 25.1%. The odds ratio for Estimated Salary, although close to 1 due to the small coefficient size, still indicates a positive association with purchase likelihood, reinforcing the significance of both predictors.

Table 4. Logistic Regression Results with Odds Ratios.

Variable	Odds Ratio	Standard Error	z-value	P>|z|	95% CI
Age	1.25094	0.03213	8.72	0.000	**[1.18953, 1.31553]**
Estimated Salary	1.000035	5.52e-06	6.32	0.000	**[1.000024, 1.000046]**
Intercept	6.28e-06	4.98e-07	-9.27	0.000	**[4.98e-07, 7.91e-07]**

Model Statistics
Log likelihood -131.3114
LR chi² (2) 215.06
Pseudo R² 0.4502
Observations 365

To quantify the practical impact of these predictors on purchase probability, we calculated the marginal effects, as shown in Table 5. For Age, the marginal effect was 0.0258, indicating that each additional year increases the probability of purchase by approximately 2.58 percentage points, holding all other variables constant. In contrast, Estimated Salary exhibited a much smaller marginal effect of 4.02e-06, meaning that for each additional dollar in estimated salary, the probability of purchase increases by just 0.0004 percentage

points. Although this effect is statistically significant, its magnitude is considerably smaller than that of Age. This suggests that while estimated salary does have a positive influence on purchase likelihood, it has a relatively minimal practical impact compared to Age. This difference implies that age plays a more substantial role in predicting purchase behavior within this model, while estimated salary serves as a secondary predictor with a subtler effect on the outcome.

Table 5. Marginal Effects of Explanatory Variables.

Variable	Marginal Effect (dy/dx)	Standard Error	z-value	P>|z|	95% CI
Age	0.0258	0.0014	18.21	0.000	**[0.0230, 0.0286]**
Estimated Salary	**4.02e-06**	**4.87e-07**	**8.25**	**0.000**	**[3.06e-06, 4.97e-06]**

Prediction Results and Model Evaluation

To assess the logistic regression model's predictive performance, the dataset was split into training and testing sets. Using a probabilistic threshold, individuals were classified as likely purchasers (1) or non-purchasers (0) based on predicted probabilities. The trained model was then applied to the test set, and its effectiveness was evaluated using a confusion matrix and metrics of accuracy, precision, and recall.

The confusion matrix in Table 6 compares model predictions to actual outcomes. Of the 35 test set observations, the model correctly classified all 24 non-purchasers (Predicted = 0), with no false positives. For the 11 actual purchasers, 8 were accurately predicted as purchasers (Predicted = 1), while 3 were misclassified as non-purchasers, resulting in false negatives.

Three main metrics provide a comprehensive evaluation of model performance: accuracy, precision, and recall. The model achieved an accuracy of 91.43%, correctly predicting purchase behavior in over 91% of cases. Precision, calculated as the proportion of true positives among all positive predictions, was 1.0000, indicating perfect reliability in predicting purchasers. Recall, measuring the model’s ability to capture actual purchasers, was 72.73%, reflecting its effectiveness in identifying most, though not all, purchasers in the test set. Table 6 summarizes the confusion matrix and evaluation metrics.

Table 6. Confusion Matrix of Model Predictions and Model Evaluation Metrics.

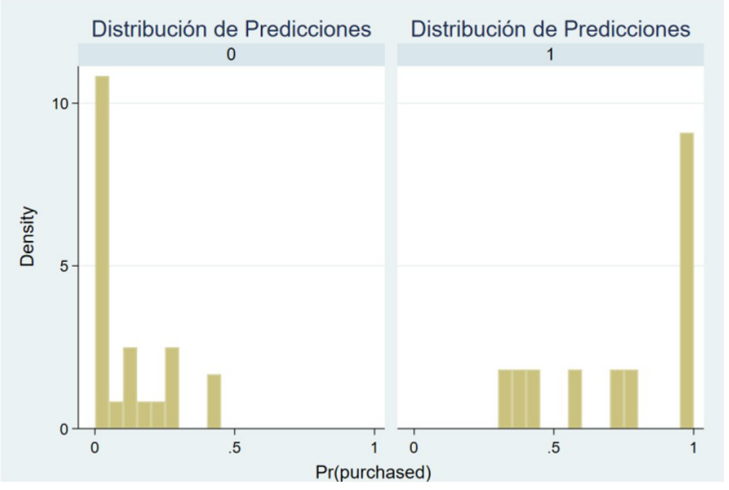
Actual Purchased	Predicted = 0	Predicted = 1	Total
0 (Not Purchased)	24	0	**24**
1 (Purchased)	3	8	**11**
Total	27	8	**35**

Metric	Value
Accuracy	0.9143
Precision	1.0000
Recall	0.7273

To further evaluate the model's predictive behavior, we analyzed the distribution of predicted probabilities for each class using the histogram in Figure 4. In the left panel (labeled "0"), representing non-purchasers, the distribution is strongly skewed towards zero, indicating high model confidence in correctly identifying non-purchasers. This aligns with the high precision in the confusion matrix, where all true negatives were accurately classified without false positives.

In the right panel (labeled "1"), representing purchasers, probabilities cluster near one, suggesting similar confidence in positive predictions. However, a few cases display lower probabilities closer to 0.5, potentially explaining the false negatives observed in the recall metric. This distribution reveals the model's conservative approach, favoring high certainty before assigning a positive label (purchase).

Figure 4. Distribution of Predicted Probabilities by Purchase Status.



Overall, this histogram illustrates the model's capacity to make well-defined predictions for each group, with strong confidence in both high and low probability ranges. However, the presence of some lower probabilities within the purchaser group highlights the trade-off between precision and recall, as the model prioritizes accuracy over a broader capture of all potential purchasers. This conservative approach aligns with the high precision observed but indicates an area where further model tuning could improve recall without sacrificing accuracy.

CONCLUSIONS

This study provides a comprehensive analysis of demographic factors influencing purchase likelihood in the context of social media advertising. Using logistic regression, we identified age and estimated salary as significant predictors of purchase behavior, each associated with an increase in purchase probability. Age demonstrated a particularly strong effect, with each additional year correlating with a 25.1% increase in purchase odds. By contrast, the effect of estimated salary, while statistically significant, was subtler in its impact, suggesting that salary influences purchase likelihood but to a lesser degree than age. This distinction was further highlighted in the marginal effects analysis, where an incremental increase in age produced a more noticeable rise in purchase probability compared to similar changes in estimated salary. This differential effect underscores age as a more powerful predictor in our model, while estimated salary serves as an additional, but secondary, factor.

The model evaluation revealed a high accuracy rate of 91.43% and a precision of 100%, underscoring the model's strength in accurately identifying purchasers with minimal false positives. However, a recall rate of 72.73% highlighted a limitation in detecting all actual purchasers, indicating that the model may be conservative in its predictions, prioritizing certainty over inclusivity. This pattern was reflected in the predicted probability distribution, which demonstrated high confidence, especially for non-purchasers.

These findings have significant implications for targeted marketing strategies. By leveraging insights into age and estimated salary as key demographic factors, marketers can improve the efficiency of ad targeting, directing resources toward segments with a higher likelihood of purchasing. Understanding these nuanced demographic influences enables companies to fine-tune their social media advertising strategies, maximizing engagement and conversion rates. This research contributes both to academic discussions on consumer behavior in digital contexts and to practical applications in optimizing advertising strategies. Future research could explore additional demographic or psychographic variables to improve model recall and achieve a balanced approach that enhances both prediction accuracy and inclusivity.

REFERENCES

Akar, E., & Topçu, B. (2011). An Examination of the Factors Influencing Consumers' Attitudes Toward Social Media Marketing. Journal Of Internet Commerce, 10(1), 35-67. https://doi.org/10.1080/15332861.2011.558456

Chaffey, D., & Ellis-Chadwick, F. (2019). Digital Marketing: Strategy and Implementation. Pearson Education.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3rd ed.). John Wiley & Sons.

Kaplan, A. M., & Haenlein, M. (2009). Users of the world, unite! The challenges and opportunities of social media. Business Horizons, 53(1), 59-68. https://doi.org/10.1016/j.bushor.2009.09.003

Kotler, P., & Keller, K. L. (2016). Marketing Management (15th ed.). Pearson.

Menard, S. (2010). Logistic Regression: From Introductory to Advanced Concepts and Applications. https://doi.org/10.4135/9781483348964

Naem, M., & Okafor, S. (2019). User-Generated Content and Consumer Brand Engagement. En Advances in marketing, customer relationship management, and e-services book series (pp. 193-220). https://doi.org/10.4018/978-1-5225-7344-9.ch009

Peng, C. J., Lee, K. L., & Ingersoll, G. M. (2002). An Introduction to Logistic Regression Analysis and Reporting. The Journal Of Educational Research, 96(1), 3-14. https://doi.org/10.1080/00220670209598786

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2017). Social media analytics – Challenges in topic discovery, data collection, and data preparation. International Journal Of Information Management, 39, 156-168. https://doi.org/10.1016/j.ijinfomgt.2017.12.002