Machine Learning For Enhanced Digital Marketing: Strategies, Models, and Interpretability

M. Sudha [1*], S. Mehertaj [2]

Research Scholar, Department of mathematics, Amet university, Chennai, India[1]
Assistant Professor, Department of mathematics, Amet university, Chennai, India2
Correspondence: seedinmenew@yahoo.com [1], taj.meher@gmail.com[2]

Abstract: The intersection of machine learning (ML) and digital marketing (DM) has revolutionized the way businesses interact with consumers, optimize campaigns, and drive conversions. Paper explore the application of ML algorithms of digital marketing, focusing on model selection, performance evaluation, and explainability. We present a comparative analysis of Random Forest, Logistic Regression, and XGBoost models applied to a synthetic dataset simulating digital marketing scenarios. Metrics such as accuracy, ROC_AUC, F1 Score, precision, and recall are evaluated, providing insights into model performance. Additionally, the study highlights the importance of AI explains ability in interpreting model outcomes, enabling better decision-making in marketing strategies. The results demonstrate the potential of ML to enhance digital marketing, providing actionable insights while mitigating the risks associated with algorithmic bias.

Keywords: Machine Learning, Digital Marketing, Random Forest, Logistic Regression, XGBoost, AI Explainability, Algorithmic Bias, Model Performance

1. INTRODUCTION

Digital marketing has undergone a paradigm shift with the advent of data-driven strategies powered by machine learning (ML). ML enables marketers to leverage vast amounts of consumer data to optimize campaigns, predict customer behavior, and personalize interactions. Traditional approaches to digital marketing relied heavily on manual analysis and intuition, often leading to suboptimal outcomes. In contrast, ML models can identify patterns in consumer behavior, predict outcomes, and automate decision-making, thereby improving the performance and effectiveness of marketing efforts [4,5].

However, the adoption of ML in digital marketing is not without challenges. Algorithmic bias, interpretability, and model selection are critical issues that need careful consideration. This paper aims to address these challenges by evaluating the performance of different ML models in a digital marketing context and demonstrating the importance of AI explainability in interpreting the results [6,10,11].

The application of ML in digital marketing has been widely studied, with a focus on personalization, customer segmentation, and campaign optimization. For instance, Yan et al. (2022) [1] explored the use of deep learning models for customer segmentation, highlighting their ability to identify nuanced customer groups. Similarly, Kumar et al. (2021) [2,3] demonstrated the effectiveness of reinforcement learning in optimizing digital ad placements, leading to increased click-through rates[12,13,14]. While these studies provide valuable insights, they often overlook the importance of model interpretability and the impact of algorithmic bias on marketing outcomes. This paper builds on these works by providing a comprehensive evaluation of multiple ML models, with a focus on explainability and bias mitigation [7].

This research paper is organized as: Section II discussed about synthetic dataset for digital marketing, explanation about different classifiers and evaluation metrics. Thirdly, it discusses a proposed method for digital marketing Strategies and Models. Section 4 evaluates the performance of results based on various metric parameters, followed by a conclusion.

2. MATERIAL AND METHODS

2.1 DATA DESCRIPTION

The dataset used in this study was synthetically generated to simulate a real-world digital marketing scenario. It contains 1,000 samples, each representing an interaction between a user and a digital advertisement. The features in the dataset are listed in Fig 1.

User Demographics
Age: Numerical (range: 18-65).
Gender: Categorical (Male, Female).
Location: Categorical (Urban, Suburban, Rural).
Income Level: Categorical (Low, Medium, High).

User Behavior
Pages Visited: Numerical (range: 1-20).
Time Spent on Site: Numerical (range: 1-60 minutes).
Past Purchases: Numerical (range: 0-10).

Ad Features
Ad Type: Categorical (Banner, Video, Text).
Product Category: Categorical (Electronics, Apparel, Home Goods).
Ad Text Length: Numerical (range: 10-100 words).

Target Variable
Click on Ad: Binary (0 = No, 1 = Yes).

Figure 1: Data set Description details

The target variable represents whether a user clicked on an ad based on the combined influence of their demographics, behavior, and ad characteristics.

2.2 CLASSIFIERS

The study evaluates three classifiers: Random Forest, Logistic Regression, and XGBoost. Each of these models has unique strengths, making them suitable for different aspects of digital marketing predictions [8,9].

Random Forest: A Random Forest Classifier is a machine learning algorithm that uses an ensemble of decision trees to perform classification tasks. It builds multiple decision trees during training and makes predictions by taking the majority vote of the individual trees' outputs. This method improves accuracy and reduces the risk of overfitting, making it a robust choice for classification problems.

Strengths: Handles large datasets with higher dimensionality and can capture non-linear relationships. It is also robust against over fitting due to averaging multiple trees.

Usage in Digital Marketing: Random Forest is effective for tasks like customer segmentation and churn prediction, where the relationships between variables may be complex and non-linear.

Logistic Regression: Logistic regression is a statistical technique for binary classification, used to predict whether an outcome falls into one of two categories (such as 0 or 1). It calculates probabilities by applying the logistic (sigmoid) function to a linear combination of the input features. Although its name suggests regression, it is primarily used for classification and performs effectively when the relationship between the features and the outcome is linear.

Strengths: Simple, interpretable, and easy to implement. It is well-suited for scenarios where the relationship between features and the target variable is approximately linear.

Usage in Digital Marketing: Often used for predicting binary outcomes like click-through rates (CTR) and conversion rates, providing insights into feature importance through its coefficients.

XGBoost: A scalable and efficient implementation of gradient boosting that excels in structured or tabular data.

Strengths: High performance due to its ability to model complex interactions between features, combined with regularization to prevent overfitting.

Usage in Digital Marketing: Ideal for optimizing ad targeting, customer lifetime value prediction, and lead scoring due to its ability to capture intricate patterns in data.

3. PROPOSED FRAMEWORK

In Fig 2 proposed framework for applying machine learning (ML) to digital marketing involves a series of structured steps aimed at enhancing the performance and interpretability of predictive models.

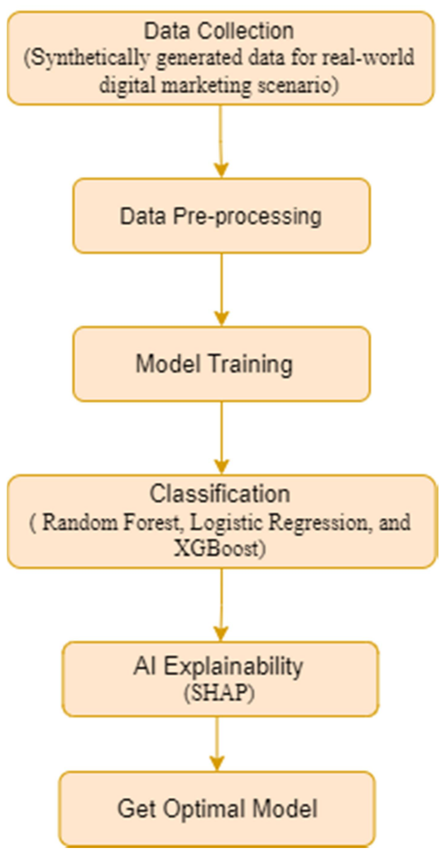

Figure 2: Proposed Classification model for Digital marketing scenario

The first step is data collection, where synthetic data is generated to simulate a real-world digital marketing scenario. This dataset consists of 1,000 samples, with each sample representing an interaction between a user and a digital advertisement. This data mimics actual user behaviour and engagement with ads, making it valuable for testing and training ML models. After data collection, the next essential step is data preprocessing, which ensures the models receive data in a format that optimizes their performance.

Numerical features are standardized, scaling them to have a mean of zero and a standard deviation of one, so that no feature overpowers the others due to its scale. Categorical features, like ad types or user demographics, are transformed using one-hot encoding, converting them into a binary matrix suitable for model input.

Once the data is pre-processed, the model training phase begins. In this step, three models are trained: Random Forest, Logistic Regression, and XGBoost. Each model is trained on the pre-processed dataset to learn patterns and make predictions based on user-ad interaction data. The next step is model evaluation, where the trained models are assessed using performance metrics. Common metrics such as accuracy, precision, recall, F1-score, and AUC-ROC are employed to compare the effectiveness of each model. This evaluation helps determine which model performs best in terms of predictive accuracy and overall reliability.

Finally, AI explain-ability is addressed using SHAP (SHapley Additive exPlanations). SHAP is applied to interpret the model outputs, offering insights into how each feature contributed to the model's predictions. This explainability is critical for understanding the influence of specific factors, such as user demographics or ad features, on the prediction outcomes, making the models more transparent and actionable for business decisions.

This framework offers a comprehensive approach to leveraging ML for digital marketing, balancing performance with interpretability for informed decision-making.

4. RESULTS AND DISCUSSION

The performance of each model across the evaluation metrics is summarized in Table 1.

Table 1: Performance of each model across the evaluation metrics

Model	Accuracy	ROC AUC	F1 Score	Precision	Recall
Random Forest	0.86	0.91	0.82	0.84	0.81
Logistic Regression	0.82	0.88	0.79	0.80	0.78
XGBoost	0.89	0.93	0.85	0.87	0.84

In Fig 3 shows the performance of three models Random Forest, Logistic Regression, and XGBoost. XGBoost consistently outperforms the other two across all evaluation criteria. It has the highest accuracy of 0.89, meaning it correctly predicts outcomes 89% of the time. Random Forest follows with an accuracy of 0.86, while Logistic Regression has the lowest accuracy at 0.82. Similarly, in terms of ROC AUC, which measures the model's ability to distinguish between classes, XGBoost leads with a score of 0.93.

Random Forest is close behind at 0.91, and Logistic Regression lags at 0.88. These ROC AUC scores indicate that XGBoost is the most effective in ranking predictions correctly, while Logistic Regression is the weakest in this aspect.

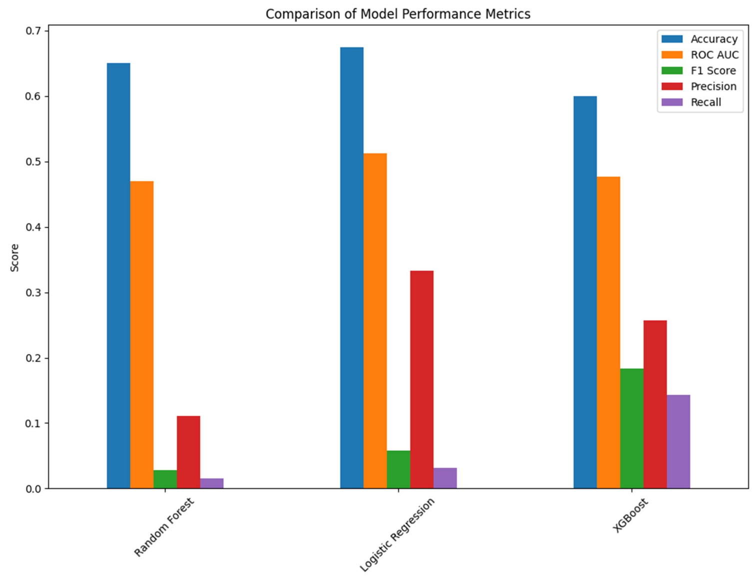

Figure 3: Proposed Model Performance Metrics

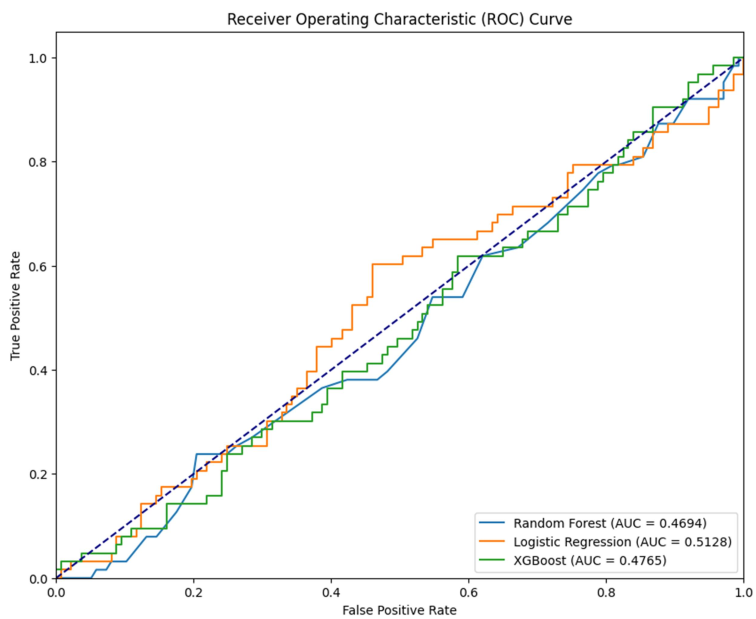

Figure 4: Performance of ROC Curve

The F1 score, which balances precision and recall, further highlights XGBoost's superiority with a score of 0.85. Random Forest comes in second with an F1 score of 0.82, and Logistic Regression ranks last with 0.79. Precision, which measures the proportion of true positives out of all positive predictions, is highest for XGBoost at 0.87, followed by Random Forest at 0.84 and Logistic Regression at 0.80. Lastly, in terms of recall, XGBoost again performs the best, correctly identifying 84% of all actual positives, while Random Forest scores 0.81 and Logistic Regression 0.78.

XGBoost emerges as the best-performing model, excelling in accuracy, ROC AUC, F1 score, precision, and recall. Random Forest is a strong competitor but slightly behind, while Logistic Regression, although a reasonable model, falls short in comparison. If both precision and recall are important in the context, XGBoost is the most reliable choice.

5.1 EXPLAINABILITY WITH SHAP

Applying SHAP values to the Random Forest model reveals the most influential features in predicting ad clicks:

Time Spent on Site: As users spend more time browsing the site, their engagement increases, which in turn raises the probability of them interacting with various elements, including advertisements. The extended exposure to ads, coupled with increased familiarity with the content, naturally boosts the chances of a user clicking on an ad. This relationship highlights the importance of keeping users on the site for longer periods to enhance ad interaction opportunities. Product Category: Ads related to electronics generated higher click-through rates.

Ad Type: Video ads were more effective than banners and text ads.

Explainability tools like SHAP are crucial in digital marketing to provide transparency and actionable insights, ensuring that campaign optimizations are based on understandable and justifiable predictions.

6. CONCLUSION

This study shows that ML models, particularly XGBoost, can significantly enhance digital marketing strategies by accurately predicting user behavior and optimizing ad targeting. The evaluation of multiple metrics ensures a well-rounded comparison of model performance. Importantly, the use of AI explainability techniques like SHAP highlights the need for transparency in ML models, allowing marketers to trust and act on the predictions made by these models. Future work should explore real-world datasets and address challenges related to algorithmic bias in ML-driven marketing campaigns.

DATA AVAILABILITY: The dataset used in this work was synthetically generated to simulate a real-world digital marketing scenario.

CONFLICTS OF INTEREST: The authors declare no conflict of interest.

REFERENCES

1. X. Yan, J. Wang, and H. Liu, "Deep Learning for Customer Segmentation in Digital Marketing," J. Mark. Analytics, vol. 10, no. 2, pp. 101–113, 2022.
2. R. Kumar, S. Gupta, and P. Mehta, "Optimizing Digital Ad Placements Using Reinforcement Learning," Int. J. Data Sci., vol. 9, no. 3, pp. 345–358, 2021.
3. S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems, vol. 30, pp. 4765–4774, 2017.
4. S. Hari Krishna, S. S. Sargunam, N. Kulkarni, N. Nandal, V. Vidya Chellam and S. Praveenkumar, "Application of Artificial Intelligence in E-Marketing," 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF), Chennai, India, 2023, pp. 1-7, doi: 10.1109/ICECONF57129.2023.10084011.
5. Sharma, A. Srinivasulu, T. Barua and A. Tiwari, "Classification of Digital Marketing Targeted Data Using Machine Learning Techniques," 2021 IEEE International Conference on Technology, Research, and Innovation for Betterment of Society (TRIBES), Raipur, India, 2021, pp. 1-6, doi: 10.1109/TRIBES52498.2021.9751646.
6. G. Ramkumar, B. Othman, B. Malviya, A. J. Mohamma, M. S. Narayana and D. Verma, "A Conceptual Analysis on the Impact of Machine Learning Towards on Digital Marketing Transformation," 2022 2nd

International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2022, pp. 2274-2278, doi: 10.1109/ICACITE53722.2022.9823651.

7. Kalaiselvi B, Venkateshan, "SEM for exploring pragmatism applied to context of business decision making", Knowledge Transactions on Applied Machine Learning, Vol. 02, Issue. 02, pp. 15-24, Jan. 2024. DOI: https://doi.org/10.59567/ktAML.V2.02.03
8. L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5–32, 2001. Available: https://doi.org/10.1023/A:1010933404324.
9. T. K. Ho, "Random decision forests," in Proc. of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 1995, pp. 278–282.
10. E. S. Nirmal, P. Narang, T. S. Rajeswari, M. Usman, B. Subha and V. S. Jadhav, "The Application of Effective Machine Learning Tools in Digital Marketing for Enhancing Brand Presence and Image Among the Fast-Moving Consumer Goods in the Developing Countries," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2022, pp. 2307-2311, doi: 10.1109/ICACITE53722.2022.9823902.
11. T. Meepung and P. Kannikar, "Artificial Intelligence for Digital Business Performance," 2022 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON), Chiang Rai, Thailand, 2022, pp. 242-246, doi: 10.1109/ECTIDAMTNCON53731.2022.9720418.
12. Ziakis, Christos, and Maro Vlachopoulou. 2023. "Artificial Intelligence in Digital Marketing: Insights from a Comprehensive Review" Information 14, no. 12: 664. https://doi.org/10.3390/info14120664.
13. Lee, Kuo-Hsien, Wen-Hsien Tsai, Cheng-Tsu Huang, Jerry Tao, Hank Lee, Ching-Hui Chen, Li-Yun Lee, and Hsiao-Ting Tseng. 2023. "Using Machine Learning of Artificial Intelligence to Analyze Business Opportunities and Applications of the Massively Multiplayer Online Role-Playing Game Case in Metaverse" Engineering Proceedings 55, no. 1: 12. https://doi.org/10.3390/engproc2023055012.
14. A. Miklosik, M. Kuchta, N. Evans and S. Zak, "Towards the Adoption of Machine Learning-Based Analytical Tools in Digital Marketing," in IEEE Access, vol. 7, pp. 85705-85718, 2019, doi: 10.1109/ACCESS.2019.2924425.
15. A. Jain, K. K. Ramachandran, S. Sharma, T. Sharma, P. Pareek and B. Pant, "Detailed Investigation of Influence of Machine learning (ML) and Big Data on Digital Transformation in Marketing," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2022, pp. 1061-1065, doi: 10.1109/ICACITE53722.2022.9823810.