# A Data-Driven Analysis and Optimization of Social Media Marketing Campaigns

***Abstract*—Social media marketing has become ubiquitous as rapidly growing user activity drives greater complexity. Despite progress in predicting advertising performances, finding actionable insights are still challenging due to the inadequate approaches for ROI(Return on Investment) maximization and the availability of comprehensive and realistic datasets still limited. this paper integrates exploratory data analysis, while leveraging feature engineering to implement Random Forest Regressor model which evaluates the campaign level ROI. A scenario-based what-if analysis is employed to assess ROI which provides decision-oriented insights and translating optimal configurations into interpretable business rules. Ultimately, the research contributes toward a structured and explainable approach for ROI analysis with actionable decision support for optimizing social media marketing campaigns.**

***Index Terms*—Social media marketing, Random forest, ROI, Feature importance, What-if analysis**

## I. Introduction

Social media marketing allows businesses to reach diverse audiences on platforms like Facebook, Instagram, X etc. Promotion of businesses mostly lies on social media nowadays as it has more active users than any other platform. But the growing amount of user data and the diversity of the audience's behavior makes it more complex to analysis target audience.

Traditionally, social media advertisement decisions were opted by manual rules, marketer experience or simple performance metrics. Though these approaches may work for small-scale business campaigns, they struggle to cope with the scale, variability and dynamic characteristics of modern advertising environment. Factors such as platform selection, target audience characteristics, campaign duration, and user engagement interaction make it difficult to identify optimal campaign strategy for advertisers. Therefore, advertising optimization has gradually shifted towards data driven approaches to increase advertising efficiency and return on investment[1].

Machine learning techniques have been widely adopted to analyze vast user data to predict customer purchase preferences and support automated decision-making using performance indicators such as click-through rate, conversion rate, and return on investment[2].Most existing approaches focus on accuracy and limited decision-making support without explaining how the advertisers should adjust configuration to improve outcomes.Similarly, most hybrid models operate as a black-box model, making it difficult for marketers to believe and trust the decisions suggested by these models[3]. This lack of interpretability reduces the practical usage of machine learning solutions in real-world advertisement.

This study aims to develop a framework that helps in decision making by using Random Forest based regressor model and feature engineering. The proposed approach predicts campaign success and provides actionable recommendations for optimization.

## II. Related Works

Social media marketing has become a highly competitive, data-driven ecosystem requiring intelligent systems to optimize campaign configurations and maximize return on investment. Rapid expansion of social media has made target group selection methods more complex because of excessive data. As a result, ad optimization has emerged as an effective approach to efficiently deliver advertisements and boost product sales. Traditional heuristic-based approaches are no longer sufficient to handle the complexity of modern advertising ecosystems, motivating the adoption of machine learning techniques for prediction, targeting, and optimization of advertising performance [3].

Previous studies showed that advertising optimization primarily focused on budget allocation and return on investment maximization. A notable contribution proposed a unified framework for marketing budget allocation by combining interpretable demand models with neural networks, enabling scalable optimization under ROI constraints. Zhao et al. [1] demonstrated significant improvements in sales while simultaneously reducing marketing costs, highlighting the effectiveness of data-driven budget allocation in large-scale advertising systems. However, this framework concentrates mainly on budget distribution and does not address campaign design variables such as platform choice, audience targeting, or campaign duration.

Beyond budget allocation, several studies applied machine learning techniques to predict consumer behavior and advertising effectiveness in social media contexts. A machine learning–based approach to enhance social media marketing employed multiple classifiers to analyze user-generated content and predict purchase behavior. The results showed that decision tree–based models achieved high prediction accuracy, confirming the suitability of supervised learning for modeling consumer response [2]. Nevertheless, this work focuses primarily on prediction and does not translate predictive outcomes into actionable campaign optimization strategies.

Targeted advertising has recently been explored using machine learning for classification and behavioral modeling. A comprehensive survey of machine learning techniques for targeted advertising categorized existing approaches into user-centric and content-centric strategies, where behavioral targeting and click-through rate prediction are the most widely

TABLE I: Summary of Related Works

| Ref. | Year | Dataset | Method | Result | Limitations |
|---|---|---|---|---|---|
| [1] | 2019 | Large-scale e-commerce advertising data (Alibaba) | Semi-black-box model (Logit + Neural Network) for budget optimization | Improved sales by >6% while reducing marketing cost by ∼40% | Focuses on budget allocation only; lacks audience, platform, and campaign design optimization |
| [2] | 2020 | Social media user-generated content | ML classifiers (J48, Naïve Bayes, SMO) | Decision tree achieved ≈96.7% accuracy in purchase behavior prediction | Prediction-only; no campaign optimization or decision support |
| [3] | 2020 | Online advertising datasets | ML taxonomy (behavioral targeting, CTR prediction) | Identified behavioral targeting and CTR prediction as core advertising techniques | Survey-based; no implementation of optimization framework |
| [4] | 2022 | Survey data (social media users) | SEM + K-means clustering | Social media marketing explains 84.1% of variance in purchase behavior | Engagement analyzed but not used for optimization or recommendations |
| [5] | 2024 | Amazon Ads campaign dataset | ML models with probabilistic modeling | Low-CPC keywords yield higher profitability; probabilistic ROI estimation | Platform-specific (Amazon); limited interpretability for campaign configuration |
| [6] | 2025 | Survey data (252 users) | LSTM | Achieved 85% testing accuracy for purchase intention prediction | Black-box model; no platform, duration, or audience optimization |
| [7] | 2025 | Social media ad campaign dataset | Ensemble learning (Random Forest, Gradient Boosting, Stacking) | Stacking achieved lowest MSE and highest $R^2$ for ROI prediction | Emphasizes prediction accuracy; lacks prescriptive decision rules |

adopted methods. While this work provides a structured overview of the field, it does not propose an integrated optimization framework that converts predictions into practical decision-making tools [4].

A study combining structural equation modeling (SEM) with unsupervised machine learning found that social network marketing activities explain a substantial proportion of variance in consumer purchase behavior. Ebrahimi et al. [7] reported that engagement-related factors such as entertainment and interaction have the strongest influence on purchase decisions, while clustering techniques revealed distinct consumer segments with different response levels. Despite these insights, engagement metrics were analyzed and were not incorporated into a prescriptive optimization framework.

Recent research has extended advertising optimization toward campaign-level performance and profitability modeling. A machine learning–based optimization framework for e-commerce advertising campaigns employed regression models, clustering, and probabilistic analysis to study the relationship between bidding variables and profitability metrics such as ACOS and ROAS. Studies showed that low-CPC keywords consistently yield higher profitability and that probabilistic modeling can support risk-aware decision-making. However, the framework promotes platform-specific guidance for optimizing broader campaign design variables in social media settings [6].

Deep learning models, such as LSTM-based approaches, have also been used to analyze the impact of functional and emotional advertising attributes on purchase intention and achieved high prediction accuracy. The results confirmed the effectiveness of deep learning for capturing temporal dependencies in consumer behavior [5]. Nonetheless, these approaches function as black-box models and focus solely on prediction, without providing explainable insights or actionable recommendations for campaign configuration.

More recently, ensemble-based approaches have been proposed to improve social media advertising performance. Studies using Random Forest, Gradient Boosting, and stacking techniques demonstrated that ensemble models outperform individual learners in predicting key advertising metrics such as ROI and CTR. The findings highlight the effectiveness and predictive strength of ensemble learning for social media advertising data [8]. However, similar to earlier works, the emphasis remains on prediction accuracy rather than decision-oriented optimization.

In summary, existing literature establishes the effectiveness of machine learning for advertising prediction, consumer engagement analysis, and isolated optimization tasks such as budget allocation or bidding. However, most studies stop at predictive modeling and fail to integrate prediction outputs into a closed-loop, explainable optimization framework that supports competitive campaign decisions such as platform selection, audience targeting, and campaign duration. These limitations motivate the development of a decision-oriented, interpretable machine learning approach that transforms predictive insights into actionable recommendations for social media advertisement.

## III. Methodology

This study proposes a structured analytical framework to analyze and predict the performance of digital marketing campaigns using statistical analysis and machine learning techniques. The methodology is made up of four parts: i) Data collection, ii) Dataset preprocessing, iii) Exploratory data analysis (EDA), iv) Predictive modeling , v) Optimization

### A. Data collection

The dataset for this study was collected from the open-source repository kaggle. It is a large scale digital marketing

campaign dataset containing 200,000 campaign records.

### B. Data Preprocessing

To ensure consistency and standardization across the dataset, preprocessing steps were applied prior to training. These included:

1) Data Cleaning: Missing values and duplicate records were checked and none were found. Then all catagorical values were standardized.Then necessary data type conversion was done
2) Feature Engineering: To enhance analytical insights, two additional features were derived:
   Cost per Click (CPC) = Acquisition Cost / Clicks
   Engagement per Impression = Engagement Score / Impressions
3) Feature Selection: Attributes that do not contribute directly to predictive performance were removed. These include Campaign ID, Company, Location, Language, and Customer Segment.
4) Encoding of Categorical Variables: Categorical attributes such as campaign type, target audience, and marketing channel were transformed using label encoding to enable their use in machine learning models.

### C. Exploratory data analysis (EDA)

Exploratory Data Analysis (EDA) was conducted to understand the distribution of variables and the relationships among campaign characteristics, engagement metrics, and performance indicators. Statistical summaries and visualizations were used to identify patterns and trends across different campaign configurations.

### D. Predictive Modeling

To predict campaign performance, a Linear Regression model was used as a baseline, followed by a Random Forest regression model to capture non-linear relationships and feature interactions.

### E. Optimization

Based on model predictions and feature importance analysis, a scenario-based optimization approach was applied to identify effective campaign configurations..

## IV. ANALYSIS & RESULTS

### A. Exploratory data analysis (EDA)

Figure 1 illustrates the average Return on Investment (ROI) across different marketing channels. The results show that Facebook and Website channels achieve marginally higher average ROI compared to other channels such as Instagram and YouTube. Although the overall variation is small, this indicates that channel selection still has a measurable impact on campaign profitability. These subtle differences justify further predictive modeling and optimization rather than relying solely on raw averages.

Figure 2 presents a heatmap of average conversion rates segmented by target audience and marketing channel. The

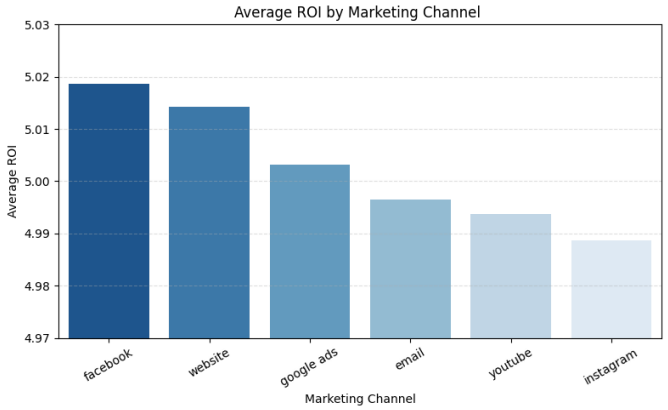


Fig. 1: Average ROI by Marketing Channel

visualization highlights slight but consistent interaction effects between audience groups and channels. For example, Google Ads and Email tend to perform marginally better for younger audiences, while Facebook and Website show relatively stable performance across age groups. This figure demonstrates that conversion behavior is influenced by the joint effect of audience demographics and channel choice, supporting the need for multi-factor optimization.

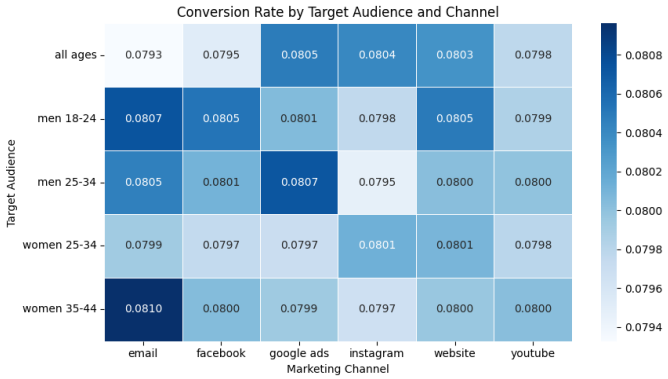


Fig. 2: Conversion Rate by Target Audience and Marketing Channel

### B. Predictive Performance Evaluation

To assess the feasibility of predicting campaign level ROI, a linear baseline model and Random Forest model were evaluated. The linear model has near zero test set $R^2$, indicating the absence of strong linear relationships between campaign attributes and ROI. The Random Forest model achieves lower

TABLE II: Predictive Performance Comparison

| Model | Train $R^2$ | Test $R^2$ | MAE | RMSE |
|---|---|---|---|---|
| Linear Regression | 0.0 | 0.0 | 1.50 | 1.74 |
| Random Forest | 0.71 | -0.02 | 0.33 | 0.39 |

MAE and RMSE compared to the linear baseline, suggesting the reduction in absolute error. Although the performance of

the test set remains close to zero (-0.02) , the model exhibits learning capacity in the training set (0.71), indicating its ability to capture non linear interactions within the data. So,the Random Forest is selected as the core model.The comparative analysis result is summarized in the given Table II.

### C. Feature Importance & What-If Analysis

The importance of features derived from the trained model demonstrated in Figure 3. The results indicate that Acquisition_Cost, impressions, cost_per_click, engagement_per_impression and click emerges as the dominant drivers of expected ROI, while categorical descriptors such as channel used, campaign type, target audience, and duration contribute marginally.

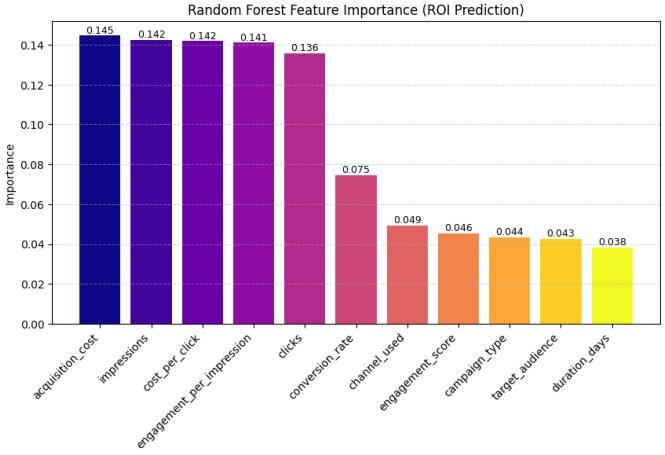


Fig. 3: Random Forest Feature Importance

TABLE III: What-If Scenario-Based Predicted ROI

| Scenario Type | Acquisition Cost | Clicks | Predicted ROI |
|---|---|---|---|
| Baseline | 12499.00 | 549 | 1.55 |
| Conservative | 8739.75 | 324 | 1.52 |
| Optimistic | 16262.00 | 774 | 1.48 |
| -10% Budget | 11249.10 | 549 | 1.54 |
| +10% Budget | 13748.90 | 549 | 1.56 |

Based on the feature distribution, a what if scenario analysis was employed to provide decision-oriented analysis that varies within high importance variables. The resulting scenario based predictions, summarized in in Table III. In particular, the baseline scenario represents a typical campaign configuration, while conservative and optimistic scenarios reflect on ROI behave under worse and better conditions. In addition, Budget intervention scenarios evaluate the directional effect of acquisition cost changes while holding other variables constant. Predicted ROI values are therefore interpreted comparatively through internal consistency, feature importance, and stability rather than exact forecasting. Under these conditions, the Random Forest model demonstrates suitability for what if analysis and decision support rather than precise outcome estimation.

### D. Recommendation Generation

Figure 4 presents the outcome of the optimization layer by visualizing the grid of rules based scenarios through a constrained decision framework.

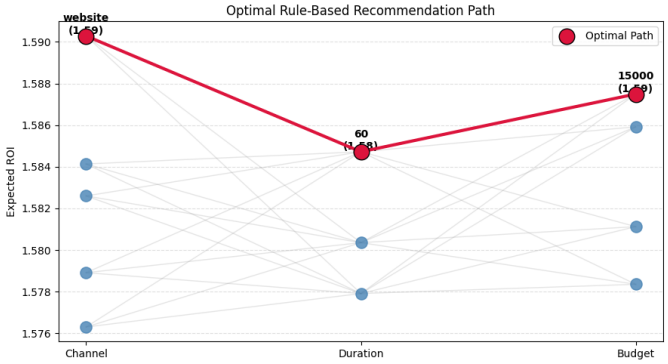


Fig. 4: Optimal ROI Recommendation Path

The decision space spans channels are Facebook, Instagram, Google Ads, and Website; campaign durations of 15, 30, 45, and 60 days; and budgets within $5k to $20k. Only feasible configurations are retained after enforcing practical constraints where **Budget** $\leq$ **$15,000**, **Engagement Score** $\geq$ **threshold**, and **Clicks** $\geq$ **historical median**. Each blue node represents a feasible decision alternative plotted against its predicted Expected ROI, while light gray paths indicate all evaluated scenario combinations generated via grid enumeration. Expected ROI is predicted for every feasible configuration, and the optimization logic selects the configuration that maximizes ROI, ensuring transparency. The red highlighted path denotes the globally optimal recommendation, identifying a Website channel, 60-day campaign duration, and $15,000 budget as the highest ROI strategy is 1.59. This optimal path is subsequently translated into actionable business rules, such as the observation that medium budgets from $12k–$15k combined with longer campaign durations yield superior ROI, thereby linking analytical optimization directly to managerial decision-making.

## V. Conclusion and Future Works

This study suggests a machine learning–based framework to analyse and improve social media marketing ROI using a dataset of 200,000 campaigns. The EDA shows that ROI varies across channels, meaning platform choice impacts performance. Predictive Performance Evaluation depicts that linear model has near zero test set R2 indicating the absence of strong linear relationships between campaign attributes and ROI. Even though performance of test set remains low in Random Forest, it has reduction in absolute error leading to choosing it for core model. A what-If scenario was introduced to predict ROI with a baseline, optimistic, conservative scenarios and also budget intervention scenarios. This shows that an increase in acquisition costs results in a higher ROI. Finally, the recommendation generation stage presents an optimization layer which searches through a limited set of campaign choices with

some given values such as keeping the budget at $15,000 limit, engagement score to minimum and click thresholds. Each configuration is evaluated using predicted expected ROIN and the framework chooses the best performing option through grid-based enumeration. The predicted outcome was converted into actionable insights, such as the concept that medium budgets along with longer durations can make better ROI which supports real marketing decision-making.

Some work that can be added to improve the whole system is to move further than direct ROI prediction and rather develop a prescriptive optimization system. The model could suggest the forward best action such as adjust budget, change channel, refine targeting , using causal inference to predict true campaign impact and avoid misleading correlations rather than only forecasting ROI. Adding more, the framework can be extended to multi-objective optimization, balancing ROI with business goals. For example conversion volume, reach, engagement or acquisition cost. On a final note, adding time-aware modelling as tracking campaign performance trends over time would allow a smarter mid-campaign intervention leading to a more realistic ROI improvement strategies.

## REFERENCES


[1] K. Zhao, J. Hua, L. Yan, Q. Zhang, H. Xu, and C. Yang, "A unified framework for marketing budget allocation," 2019. [Online]. Available: https://arxiv.org/abs/1902.01128

[2] B. Arasu, B. Seelan, and T. Natarajan, "A machine learning-based approach to enhancing social media marketing," *Computers Electrical Engineering*, vol. 86, p. 106723, 09 2020.

[3] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019. [Online]. Available: https://www.nature.com/articles/s42256-019-0048-x

[4] "Identifying machine learning techniques for classification of target advertising," *ICT Express*, vol. 6, no. 3, pp. 175–180, 2020.

[5] J. Zhu, "The impact of social media advertising on consumers' purchase intention: A study based on the lstm model," in *Proceedings of the 2025 6th International Conference on Computer Information and Big Data Applications*, ser. CIBDA '25. New York, NY, USA: Association for Computing Machinery, 2025, p. 1146–1153. [Online]. Available: https://doi.org/10.1145/3746709.3746905

[6] A. Jha, P. Sharma, R. Upmanyu, Y. Sharma, and K. Tiwari, "Machine learning-based optimization of e-commerce advertising campaigns," 01 2024, pp. 531–541.

[7] P. Ebrahimi, M. Basirat, A. Yousefi, M. Nekmahmud, A. Gholampour, and M. Fekete-Farkas, "Social networks marketing and consumer purchase behavior: The combination of sem and unsupervised machine learning approaches," *Big Data and Cognitive Computing*, vol. 6, no. 2, 2022. [Online]. Available: https://www.mdpi.com/2504-2289/6/2/35

[8] S. Turgay, M. Kavacik, Y. Tonkul, M. Şahin, and R. Güzel, "Enhancing social media ad campaigns through ensemble-based optimization," , vol. 19, pp. 47–57, 06 2025.