

Statistical Profiling of Marketing Strategy Data Using Machine Learning Models

Raja Venkat Ram V*, M Raghavender Sharma

Department of Statistics
Osmania University, Hyderabad, India

Abstract—Marketing teams collect large volumes of campaign data, yet privacy rules and commercial confidentiality often prevent that data from being shared for research. This paper presents a statistical profiling framework for marketing strategy data and demonstrates it on a synthesized dataset of 5,000 campaigns that vary by channel, customer segment, creative format, season, budget, duration, discount, and audience size. The profile combines descriptive statistics, nonparametric group comparisons, rank correlations, supervised machine learning models, and K-means segmentation. Gradient boosting predicted campaign return on investment (ROI) best on held-out data (R -squared = .74, RMSE = 2.51), ahead of random forest and linear regression, and all three classifiers separated profitable from unprofitable campaigns with accuracy near .91. Channel and customer segment explained most of the variation in ROI, while budget, duration, and discount level added little. Because the data are simulated, the findings show that the pipeline recovers the relationships built into the generator; they are not evidence about real markets. The framework, data generator, and results can be reused to prototype analytics before real data are available.

Keywords—statistical profiling; marketing analytics; machine learning; synthetic data; campaign ROI; customer segmentation

I. INTRODUCTION

Marketing decisions about where to spend, whom to target, and which message to run increasingly rest on data (Wedel & Kannan, 2016). Analysts first need to understand what the data look like: how outcomes are distributed, which factors move them, and how reliably they can be predicted. This first step is often called statistical profiling, and it determines which later modelling choices make sense (Hastie et al., 2009).

Real campaign records are hard to obtain for study. They contain customer information, reveal competitive strategy, and are governed by data protection rules. Synthetic data offer a practical alternative because they reproduce the structure of a data set without exposing real records (Jordon et al., 2022; Nikolenko, 2021). They also have a research advantage: the generating process is known, so an analysis method can be checked against relationships that are true by construction.

This paper builds a reproducible profiling pipeline for marketing strategy data and tests it on a synthesized campaign data set. It addresses three questions. First, how are campaign outcomes distributed, and which strategic factors are associated with them? Second, how accurately can machine learning models predict campaign ROI and profitability? Third, do campaigns fall into distinct performance groups? The contribution is a documented workflow that combines classical statistics with machine learning, and a worked example that others can rerun.

II. BACKGROUND

A. Marketing analytics

Marketing analytics links strategic choices such as channel mix, pricing, and timing to outcomes such as reach, response, and profit (Kotler & Keller, 2016). Wedel and Kannan (2016) review how growth in data volume and variety has moved the field from descriptive reporting toward prediction and resource allocation. Campaign ROI is a common summary outcome, but it is usually skewed, which affects the choice of statistical tests.

B. Synthetic tabular data

Synthetic data can be produced by hand-specified simulation or by learned generative models such as conditional generative adversarial networks (Xu et al., 2019). Learned generators aim to copy an existing data set. A hand-specified generator, used here, instead encodes explicit assumptions about how the market works, which makes results easy to interpret and verify.

C. Machine learning models

Three supervised learners are used. Linear regression and logistic regression serve as interpretable baselines (James et al., 2021). Random forests average many decorrelated decision trees and handle nonlinear effects and interactions without manual specification (Breiman, 2001). Gradient boosting builds trees sequentially, each correcting the errors of the previous ones (Friedman, 2001). K-means clustering (Lloyd, 1982) is used for unsupervised segmentation.

III. METHODOLOGY

A. Synthetic data design

The data set contains 5,000 simulated campaigns generated in Python with a fixed random seed of 42 (McKinney, 2010). Table 1 lists the variables. Each campaign is assigned a channel, customer segment, creative format, and quarter. Budget and audience size follow log-normal distributions, duration is uniform between 7 and 90 days, and discount takes one of seven levels from 0% to 30%. Impressions are budget divided by a channel-specific cost per thousand impressions with random variation. Clicks and conversions are drawn from binomial distributions whose rates depend on channel, creative format, discount, and quarter. Revenue equals conversions multiplied by a segment-specific order value, reduced by the discount and multiplied by random noise. ROI is (revenue – budget) / budget, and a campaign is labelled profitable when ROI exceeds zero.

Table 1
 Variables in the Synthesized Marketing Data Set

Variable	Type	Values or distribution
Channel	Categorical	Email, Social Media, Search Ads, Display, Influencer, TV
Segment	Categorical	Youth, Adults, Seniors, Premium
Creative format	Categorical	Video, Image, Text, Carousel
Quarter	Categorical	Q1–Q4 (Q4 most frequent)
Budget (USD)	Numeric	Log-normal
Duration (days)	Numeric	Uniform, 7–90
Discount (%)	Numeric	0, 5, 10, 15, 20, 25, 30
Audience size	Numeric	Log-normal
Impressions, clicks, conversions	Outcome	Simulated counts
Revenue, ROI, profitable	Outcome	Derived from the above

Note. Channel and segment parameters were chosen to resemble plausible industry patterns but are assumptions, not estimates from real firms.

B. Statistical profiling

Numeric variables are summarized with means, standard deviations, medians, and skewness. Because ROI is strongly right-skewed, differences across channels, segments, quarters, and creative formats are tested with the Kruskal–Wallis test (Kruskal & Wallis, 1952), and effect size is reported as epsilon-squared (ϵ^2), calculated as $H / (n - 1)$. Associations among numeric variables use Spearman rank correlation. Following Cohen (1988), correlations near .10, .30, and .50 are read as small, medium, and large.

C. Predictive modelling and evaluation

Campaign ROI is predicted with linear regression, random forest, and gradient boosting, and profitability is classified with logistic regression, random forest, and gradient boosting. Predictors are the four categorical variables plus log budget, duration, discount, and log audience size. Outcomes observed only after a campaign ends (impressions, clicks, conversions, revenue) are excluded to prevent leakage. The data are split 80% for training and 20% for testing, stratified by profitability. Regression models are compared with five-fold cross-validated R-squared on the training set, then scored on the test set with R-squared, root mean squared error (RMSE), and mean absolute error (MAE). Classifiers are scored with accuracy, precision, recall, F1, and area under the ROC curve (AUC). Variable importance is measured by permutation importance on the test set, with 15 repeats. All models use scikit-learn (Pedregosa et al., 2011).

D. Segmentation

K-means is applied to standardized log budget, click-through rate (CTR), conversion rate, log revenue, and capped ROI. The number of clusters, from 2 to 7, is chosen by the silhouette coefficient (Rousseeuw, 1987).

IV. RESULTS AND DISCUSSION

A. Descriptive profile

Table 2 summarizes the main variables. Revenue (skewness = 13.4) and ROI (skewness = 4.2) are strongly right-skewed: the mean ROI of 2.02 is far above the median of 0.18, because a minority of very successful campaigns pull the mean upward. About 54% of campaigns were profitable. Medians and rank-based methods are therefore more informative than means for this outcome.

Table 2
 Descriptive Statistics for Selected Variables ($N = 5,000$)

Variable	M	SD	Mdn	Skew
Budget (USD)	6259	5801	4527	3.04
Impressions (000)	433	533	259	3.82
CTR (%)	1.65	1.02	1.50	0.79
Conversion rate (%)	2.50	1.37	2.15	1.11
Revenue (USD)	18754	49270	5633	13.38
ROI	2.02	5.03	0.18	4.24

Note. CTR = click-through rate; ROI = return on investment, (revenue – budget) / budget.

B. Group differences

Channel was strongly related to ROI, $H(5) = 3783.60$, $p < .001$, $\epsilon^2 = .76$. As Table 3 shows, email ($Mdn = 5.45$) and search ads ($Mdn = 1.93$) were the strongest channels, whereas display and television had negative median ROI. Customer segment also mattered, $H(3) = 523.08$, $p < .001$, $\epsilon^2 = .10$, with premium customers highest ($Mdn = 1.77$) and youth lowest ($Mdn = -0.34$). Quarter had a statistically significant but small effect, $H(3) = 42.41$, $p < .001$, $\epsilon^2 = .01$; the fourth quarter led ($Mdn = 0.45$). Creative format showed a similarly small effect, $H(3) = 62.54$, $p < .001$, $\epsilon^2 = .01$. With 5,000 campaigns, small effects reach significance easily, so effect sizes should guide interpretation more than p values.

Table 3
 Campaign Performance by Channel

Channel	n	M ROI	Mdn ROI	CTR (%)	Profitable (%)
Email	918	8.33	5.45	2.15	100
Search Ads	1111	2.81	1.93	2.92	95
Influencer	592	0.08	-0.22	1.73	37
Social Media	1207	0.04	-0.22	1.18	35
Display	679	-0.48	-0.64	0.54	12
TV	493	-0.87	-0.90	0.43	0

Note. Channels are sorted by mean ROI. CTR = click-through rate.

C. Associations among numeric variables

Figure 1 shows Spearman correlations. ROI was strongly associated with CTR ($\rho = .79$) and conversion rate ($\rho = .75$), and the two engagement measures were themselves closely related ($\rho = .72$). Discount was positively related to conversion rate ($\rho = .23$, a small-to-medium effect) but almost unrelated to ROI ($\rho = .02$). In other words, discounts lifted conversions but also reduced revenue per sale, so they did not improve return. Budget showed essentially no association with ROI ($\rho = -.03$), which suggests that spending more did not by itself produce a better return in this data set.

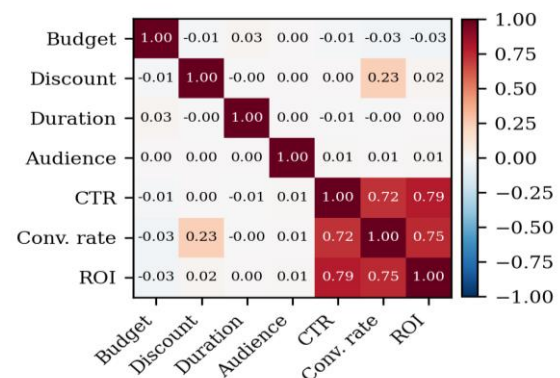


Figure 1. Spearman rank correlations among numeric campaign variables

D. Predictive performance

Table 4 compares the regression models. The two tree ensembles clearly outperformed linear regression. Gradient boosting had the best test performance (R-squared = .74, RMSE = 2.51), with random forest close behind (R-squared = .71). Linear regression explained only about 55% of the variance because the effect of channel and segment on ROI is multiplicative rather than additive. Cross-validated R-squared values agree with the test results, which indicates that the gap is not an artifact of one particular split.

Table 4
 Regression Models Predicting Campaign ROI

Model	CV R ² (M ± SD)	Test R ²	RMSE	MAE
Linear Regression	.56 ± .02	.55	3.30	1.89
Random Forest	.72 ± .05	.71	2.64	1.16
Gradient Boosting	.73 ± .03	.74	2.51	1.19

Note. CV = five-fold cross-validation on the training set; RMSE = root mean squared error; MAE = mean absolute error. Test metrics use the 20% hold-out set.

For the profitability label, the three classifiers performed almost identically (Table 5), with an accuracy of about .91 and AUC near .97. Deciding whether a campaign breaks even depends mainly on channel and segment, which a simple logistic model captures well. The extra flexibility of ensembles mattered for predicting the size of ROI but not for the yes-or-no decision.

Table 5
 Classification Models Predicting Campaign Profitability

Model	Accuracy	Precision	Recall	F1	AUC
Logistic Regression	.91	.93	.90	.91	.97
Random Forest	.91	.93	.89	.91	.97
Gradient Boosting	.91	.92	.90	.91	.97

Note. Metrics are computed on the 20% hold-out set. The profitable class is treated as positive.

E. Variable importance

Permutation importance for the gradient boosting model (Figure 2) ranks channel first (R-squared drop = 1.30) and customer segment second (0.64). Quarter contributed a small amount (.058), and creative format, budget, discount, duration, and audience size contributed essentially nothing. This ranking matches the generator, in which channel and segment set the cost and revenue structure while the other factors only modify it slightly. The agreement is reassuring about the method, but it should not be read as a claim that budget or discount are unimportant in real campaigns.

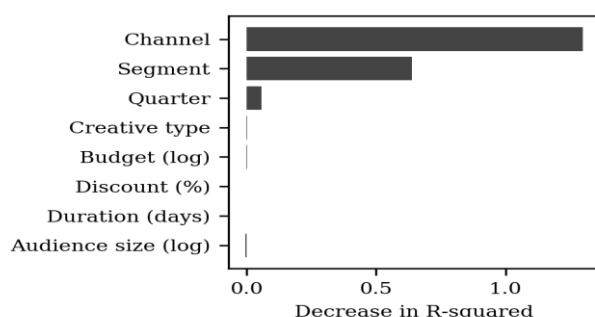


Figure 2. Permutation importance of predictors (gradient boosting model, test set)

F. Campaign segments

The silhouette coefficient was highest for two clusters (.39) and declined for larger numbers, so two segments were retained. The larger group (3,063 campaigns, 61%) had low engagement (mean CTR = 1.04%, mean conversion rate = 1.68%), a median ROI of -0.44, and only 25% profitable campaigns; its most common channel was Social Media. The smaller group (1,937 campaigns) had higher engagement (CTR = 2.61%, conversion rate = 3.79%), a median ROI of 3.47, and 99% profitable campaigns, most often through Search Ads. A silhouette value near .39 indicates moderate, not sharp, separation, so the groups are best read as a broad high-versus-low performance split.

G. Implications and limitations

For practitioners, the profile suggests an order of analysis: begin with channel and segment, use rank-based summaries for skewed outcomes, and reserve complex models for predicting the magnitude of return rather than simple profitability. For researchers, the framework offers a transparent test bed in which every true effect is known.

The main limitation is that all results come from simulated data. The strong effects of channel and segment, the near-zero effect of budget, and the discount trade-off all follow from parameters chosen by the author. The study therefore establishes that the pipeline works, not that the same patterns hold in practice. Real data would also bring missing values, delayed conversions, overlapping campaigns, and causal confounding, none of which are modelled here. Future work should repeat the analysis on anonymized real campaign records and on data from learned generators.

V. CONCLUSION

This paper presented a statistical profiling framework for marketing strategy data and applied it to a synthesized data set of 5,000 campaigns. Descriptive statistics showed highly skewed returns, rank-based tests identified channel and customer segment as the dominant factors, tree ensembles predicted ROI more accurately than linear regression, and K-means separated campaigns into high- and low-performing groups. Because the data are simulated, these results demonstrate the method rather than market facts. The workflow can be reapplied directly when real campaign data become available.

REFERENCES

- [1] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- [2] Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- [3] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232. <https://doi.org/10.1214/aos/1013203451>
- [4] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
- [5] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer.
- [6] Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., & Weller, A. (2022). *Synthetic data – what, why and how?* arXiv. <https://doi.org/10.48550/arXiv.2205.03257>
- [7] Kotler, P., & Keller, K. L. (2016). *Marketing management* (15th ed.). Pearson.

- [8] Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260), 583–621. <https://doi.org/10.1080/01621459.1952.10483441>
- [9] Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2), 129–137. <https://doi.org/10.1109/TIT.1982.1056489>
- [10] McKinney, W. (2010). Data structures for statistical computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). <https://doi.org/10.25080/Majora-92bf1922-00a>
- [11] Nikolenko, S. I. (2021). *Synthetic data for deep learning*. Springer.
- [12] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- [13] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
- [14] Wedel, M., & Kannan, P. K. (2016). Marketing analytics for data-rich environments. *Journal of Marketing*, 80(6), 97–121. <https://doi.org/10.1509/jm.15.0413>
- [15] Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. In *Advances in Neural Information Processing Systems* (Vol. 32). Curran Associates.