Journal of Data Science 0 (0), 1–14
??? 2025
DOI: 10.6339/25-JDS1198
Data Science in Action

Label-efficient Response Modelling: Cost-Effective Marketing Using Cluster-Based Active Sampling

Swee Chuan Tan[1]
[1]*School of Business, Singapore University of Social Sciences, Singapore*

Abstract

This paper introduces a label-efficient response modelling method useful when the target labels are unknown a priori. Unlike most response modelling methods that adopt a supervised or semi-supervised approach, we apply clustering to partition data into homogeneous segments, which are assumed to reflect the underlying response behaviours. We then take a random sample from each cluster. For each sampled record, the true target label is acquired. Through this cluster-based stratified sampling approach, we reduced the cost of label acquisition needed to estimate the cluster-specific and overall basic response rates. The goal is to identify a subset of the population more likely to respond (e.g., make a purchase) while controlling campaign costs. This idea of subsetting the population represents a departure from conventional classification tasks, which require full labeling of all observations. We regard clusters with response rates significantly higher than the estimated basic response rate as high-propensity clusters and proceed to acquire all their remaining labels. Our experimental results show that the response rates of high-propensity clusters are at least 1.7 times the basic response rate. This suggests that the proposed approach significantly reduces costs by targeting only high-propensity groups and is useful in scenarios lacking historical ground truth.

Keywords *active learning; data-efficient learning; imbalanced data; predictive modelling; semi-supervised learning; stratified sampling*

1 Introduction

Traditional mass marketing can help raise awareness of products or services to a broad audience. However, this approach lacks precise targeting and often results in lower conversion rates and reduced return on investment (Thomas, 2007). In addition, frequent mass outreach to low-propensity customers can lead to unintended negative brand perceptions. By focusing on high-quality leads, conversion efficiency can be improved with lower marketing costs.

To improve marketing conversion rates, one can adopt a direct marketing strategy (Housden and Thomas, 2002; Tékouabou et al., 2022) by targeting customers who are more likely to respond to campaigns or make purchases. Constructed based on historical campaign data, a predictive model (Hanssens et al., 2005) can identify potential customers along with the response probability (or propensity). By prioritizing outreach to high-propensity customers through direct marketing, significant cost savings (Gönül and Hofstede, 2006) can be achieved.

Apart from applications in marketing, predictive modelling can be adopted in many domains. For example, in healthcare (Mohammed Amine Naji et al., 2021), higher response rates can impact lives, which is invaluable, and the benefits go beyond monetary terms. Similarly, in fraud detection (Ali et al., 2022), improved response rates not only cut financial losses but also strengthen an organisation's image and reputation. Since predictive response modelling would

Received April 16, 2025; Accepted July 18, 2025

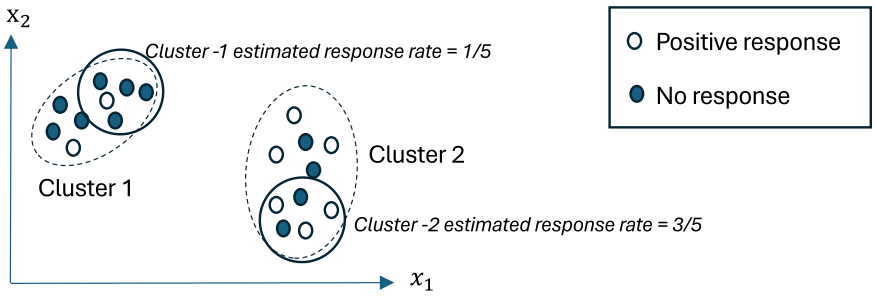


Figure 1: Estimated cluster response rates based on samples stratified by Clusters 1 and 2.

require target labels for model construction, it can be challenging when working with problems lacking historical ground truth. This problem is exacerbated in large and imbalanced datasets (He and Garcia, 2009), where target label acquisition becomes costly and model training becomes challenging.

We propose a label-efficient response modelling approach using cluster-based stratified sampling. The key idea is to group the instances into some distinct clusters, where each cluster contains similar instances. By randomly sampling a small but representative subset of instances from each cluster, only a small number of target labels are obtained within each cluster subsample. These subsamples are used to estimate the cluster-specific response rates, which can then be used to infer the basic response rate of the entire dataset. Subsequently, at 5% level of significance, clusters with response rates higher than the basic response rates are further categorised as high-propensity clusters. By targeting high-propensity clusters, we can effectively increase response rates with significantly lower costs, time, and resources.

To explain the basic idea of the proposed method, we use a toy example of customer records consisting of two input variables. The aim is to identify customers likely to respond to a marketing campaign by first performing clustering. Figure 1 shows two distinct clusters (enclosed in dotted boundaries) generated from the dataset. Particularly, Clusters 1 and 2 contain a mix of customers who may or may not respond to the marketing campaign. Instead of targeting all customers, we randomly draw a sample of five customers from each cluster (enclosed by solid circles) and engage them in the campaign. By analysing their responses, we identify high-propensity clusters that have sample cluster response rates higher than the basic response rate. For example, the estimated basic response rate in Figure 1 is about 41%, so Cluster 1 with 20% estimated response rate should be discarded; and Cluster 2 with 60% estimated response rate should be used for the subsequent marketing campaign. This method assumes each cluster contains customer records with similar profiles and propensity to take a desired action. By targeting the high-propensity clusters, this approach saves marketing costs by avoiding individuals unlikely to respond and consequently lifting the response rates. In Figure 1, the estimated response rate is lifted from 41% to 60%. If we proceed to target all 10 customers in Cluster 2, then the actual response rate is lifted from 42% to 60%.

2 Related Work

This section reviews key approaches and challenges in response modelling. In direct marketing, the primary aim of response modelling is to increase response rates (Haughton and Oulabi, 1993), where even a marginal increase can yield significant benefits. For example, Baesens (2004)

illustrated that "an increase of 1% response rate of a direct mailing campaign can result in substantial profit gains".

While the main objective of response modelling is to increase response rates, achieving this goal involves addressing several issues. First, the large volume of data in modern systems demands high computational resources for model construction. Second, the class imbalance problem is prevalent in the marketing context, where the proportion of customers responding to a marketing campaign or making a purchase represents the minority of a population. The class imbalance problem often results in biased models that favour the majority class. Third, there is often a lack of ground truth labels in cold-start scenarios (e.g., new marketing campaign), which affects the reliability of initial models.

In data-intensive systems such as e-commerce platforms, the abundance of user interaction data has prompted researchers such as Chaudhuri et al. (2021) to apply deep learning techniques for improving purchase prediction performance. While deep learning often outperforms traditional methods in complex datasets, it requires significant computational resources and infrastructure. Thus, deep learning is typically used only when the potential financial returns justify the investment in model training and deployment.

In direct marketing, Kang et al. (2012) addressed the class imbalance problem by combining clustering, under-sampling, and ensemble methods. Their approach uses clustering to identify data regions containing uneven class distributions and applies localised data balancing within these regions. However, their method assumes that true labels are available for balancing, which may not be the case in real-world scenarios. For example, class labels are usually not available or scarce in new markets.

To handle the label scarcity issue, Lee et al. (2010) adopted semi-supervised learning using transductive support vector machines (TSVMs), leveraging both labelled and unlabelled data. Their findings reveal that TSVMs improve response modelling performance even with limited labels. Apart from TSVMs, neural networks can also be used for semi-supervised learning (Emtiyaz and Keyvanpour, 2011). Initially, a classifier is constructed using labelled data, which is then used to label the remaining unlabelled data, followed by classifier refinement. Their experiments on two Customer Relationship Management datasets showed better prediction performance compared to several established methods.

Apart from semi-supervised learning, active learning is also a good option for improving label efficiency. For example, Yan et al. (2022) introduced a clustering-based active learning method to find informative and representative samples, which achieves performance comparable to state-of-the-art methods. However, active learning methods still require iterative model training, which can be computationally intensive. In addition, careful selection of the most informative data points for labelling is crucial, and a biased selection strategy often leads to suboptimal results.

In response to the challenges mentioned above, this work proposes a label-efficient and cost-effective response modelling approach. Using cluster-based active sampling (Tipton, 2013), the method first groups similar customers into distinct clusters, and within each cluster, a small subset is then randomly sampled. These cluster samples are then used to estimate cluster-specific response rates, which are aggregated — weighted by their respective cluster sizes to infer the overall response rate. Statistical testing is then applied to identify the high-propensity clusters whose response rates significantly exceed the estimated baseline. This targeted approach allows more precise and impactful customer engagement, with reduced labeling effort, computational overhead, and marketing spend, making it a practical alternative to traditional supervised or active learning methods.

3 Proposed Method

Given a dataset D with inputs X and an unknown binary target Y, where $y \in Y$ represents individual target values (0 for negative responses, 1 for positive responses), the proposed method is presented as follows.

1. **Apply K-means Clustering:** Apply K-means clustering on D. For example, $k = 2$ is the number of clusters generated in Figure 1. If k is an unknown, one can use a cluster validity index to estimate the k value. In practice, it is usually fine to choose a k value higher than what the cluster validity index suggests. This is because we do not concern ourselves with generating the optimal cluster structure for semantic interpretation. Rather, we aim to identify sub-regions in the data space that contain high-propensity customers (Haron, 2022). In this sense, the value of k need not be optimal.
2. **Determine sample size for each cluster:** For each cluster i with size n_i, determine the minimum cluster sample size n_i' using the following empirical rule:

$$n_i' = \min(n_i, n_{\max}),$$

where $n_{\max}$ is the maximum sample size for any cluster i. If n_i is greater than $n_{\max}$, then $n_i' = n_{\max}$; otherwise, $n_i' = n_i$. For the toy example in Figure 1, $n_{\max} = 5$, so $n_1' = n_2' = 5$, resulting in subsamples of size 5 in each of Clusters 1 and 2.
3. **Determine target labels and estimate response rates:** For each cluster i, determine the target label y_j of each instance j in the random sample of size n_i'. Then, estimate the sample cluster response rate for each cluster i using:

$$\hat{r}_i = \frac{\sum_{j=1}^{n_i'} y_j}{n_i'}.$$

The estimated basic response rate $\hat{r}_B$ for the entire dataset D is:

$$\hat{r}_B = \frac{\sum_{i=1}^{k} n_i \hat{r}_i}{\sum_{i=1}^{k} n_i}.$$

Note that n_i instead of n_i' must be used to estimate $\hat{r}_B$ because larger clusters have more weight on the estimated response rate, as compared to smaller clusters. Thus, the individual $\hat{r}_i$ is weighted according to the actual size n_i of its cluster i.
The actual cluster response rate is:

$$r_i = \frac{\sum_{j=1}^{n_i} y_j}{n_i}.$$

Note that r_i is not known unless all the records in cluster i are surveyed, and this would be the case if cluster i is deemed high propensity at a later stage.
The corresponding actual basic response rate is:

$$r_B = \frac{\sum_{i=1}^{k} n_i r_i}{\sum_{i=1}^{k} n_i}.$$

It is important to note that r_B is not known in practice unless the entire population is surveyed. Based on the example in Figure 1, the sample cluster response rates are $\hat{r}_1 = \frac{1}{5}$ and $\hat{r}_2 = \frac{3}{5}$. The actual cluster response rates are $r_1 = \frac{2}{9}$ and $r_2 = \frac{6}{10}$. The estimated basic response rate is $\hat{r}_B = (9 \times \frac{1}{5} + 10 \times \frac{3}{5})/19 = 0.41$, and the corresponding actual basic response rate is $r_B = \frac{8}{19} = 0.42$.

4. **Identify high-propensity clusters:** A cluster i is deemed high propensity if $\hat{r}_i > \hat{r}_B$ at the 5% level of significance. Specifically, we apply the two-proportion z-test to ascertain that the difference is statistically significant:

$$Z = \frac{\hat{r}_i - \hat{r}_B}{\sqrt{\hat{r}_B(1-\hat{r}_B)(\frac{1}{n_i'} + \frac{1}{n})}}.$$

   Note that this z-test is not suitable for the toy example in Figure 1, which has a small sample size and does not fulfill the normal approximation assumption. In practice, however, this test is useful because real-world datasets are large, allowing for more precise response rate estimates and reliable statistical inference.
5. **Estimate overall lifted response rate:** For all the high-propensity clusters, the estimated overall lifted response rate is:

$$\hat{r}_L = \frac{\sum_{i=1}^{k} n_i \hat{r}_i}{\sum_{i=1}^{k} n_i} \quad \text{s.t.} \quad \hat{r}_i > \hat{r}_B \quad \text{with} \quad p < 0.05.$$

   This enhanced response rate is achieved by targeting high-propensity clusters that are deemed significant at 5% level. It is the weighted average of response rates from high-propensity clusters, using the respective cluster sizes as weights. This metric measures the effectiveness of the proposed method in prioritizing clusters with high response rates. In general, $\hat{r}_L > \hat{r}_B$ will help optimize targeting strategies and resource allocation.
   The corresponding actual lifted response rate is:

$$r_L = \frac{\sum_{i=1}^{k} n_i r_i}{\sum_{i=1}^{k} n_i} \quad \text{s.t.} \quad \hat{r}_i > \hat{r}_B \quad \text{with} \quad p < 0.05.$$

   Note that r_L can be computed when all the records in the high-propensity clusters are surveyed. Based on the example in Figure 1, the estimated overall lifted response rate is $\hat{r}_L = \frac{3}{5}$, and the corresponding actual lifted response rate is $r_L = \frac{6}{10}$.
6. **Estimate Lift:** The lift in response rates is estimated as follows:

$$\widehat{Lift} = \frac{\hat{r}_L}{\hat{r}_B}.$$

   Similarly, the actual lift is:

$$Lift = \frac{r_L}{r_B}.$$

   Note that $Lift$ is not known in practice unless the entire population is surveyed. We include it here for the sake of performance evaluation. Based on the example in Figure 1, the estimated lift is $\widehat{Lift} = \frac{0.6}{0.41} = 1.46$, and the corresponding actual lift is $Lift = \frac{0.6}{0.42} = 1.43$.

4 Datasets, Experimental Setup, and Results

4.1 Datasets

To evaluate the proposed method, two marketing-related datasets are used. The first dataset is Bank Marketing (Moro et al., 2014), which contains 16 attributes and one binary target.

Table 1: Two imbalanced datasets used for label-efficient response modelling.

	Online shopping	Bank marketing
# Inputs	17	16
# Records	12,330	45,211
Response Rate	15.47%	11.7%
Target	{Revenue (1), No revenue (0)}	{Buy (1), No buy (0)}

Here, 11.7% of the records contain target value 'Buy' (coded as 1), and the remaining records contain target value 'No buy' (coded as 0). The second dataset, Online Shoppers Purchasing Intention (referred to as Online Shopping) (Sakar and Kastro, 2018), contains 17 attributes and a binary target. In this dataset, 15.47% of the records have the target value 'Revenue' coded as 1, indicating that a purchase was successfully completed. The remaining records, with target value 'No revenue' coded as 0, indicate web sessions without a completed transaction. These datasets are publicly available, and their details are given in Table 1. For clustering, categorical variables have been numerically encoded, and numeric variables have been standardised. Interested readers can refer to the supplementary materials for Python code details.

4.2 Parameter Settings

There are two key parameters in the proposed method. The first is the number of clusters, k. From our experience, it is generally advisable to set k above the estimated number of clusters. While this may lead to a few additional redundant clusters, these smaller clusters are usually more homogeneous. Moreover, a cluster sample size is usually much smaller than its actual cluster size (i.e., $n_i' \ll n_i$), so the additional cost of sampling redundant clusters is marginal. In general, we recommend $k \geqslant 10$. This recommendation is based on practical intuition and empirical observations. While increasing k typically refines the data partitioning and improves performance, empirical results show that performance tends to stabilise when $k \geqslant 10$. This value thus offers a good balance between model precision and the effort required for labelling.

The second parameter is $n_{\max}$, the maximum cluster sample size for any cluster. In all our experiments, we set $n_{\max} = 100$. While this is twice the typical minimum sample size of 50 required for proportion estimation, it ensures that the two-proportion z-test (for identifying high-propensity clusters) remains robust and less sensitive to the normality assumptions. Specifically, this choice satisfies the conditions

$$np \geqslant 10 \quad \text{and} \quad n(1-p) \geqslant 10$$

even for extreme proportions (e.g., $p = 0.1$). Thus $n_{\max} = 100$ provides a practical trade-off between annotation cost and statistical reliability.

4.3 Experimental Setup

Since K-means clustering may generate different results depending on the cluster centroid initialization, we repeat the clustering process 30 times and measure the average performance in our results. In addition, we evaluate the performance of the proposed method by:

- Studying the impact of varying the number of clusters on lifted response rates, where the number of clusters ranges from 2 to 30, in increments of 2.

- Studying how removing informative attributes influences model performance. In the experiments, the most informative attributes are progressively removed from each dataset, until a maximum of 12 attributes have been removed. Here, informative attributes are those whose values vary the most across cluster centres. Features with higher cross-cluster variations are better at separating data into meaningful clusters. The variations are measured using standard deviation, which aligns with the clustering assumption.
- Investigating how noisy attributes affect model performance by systematically introducing 20, 40, 60, and 80 additional noise variables. Each noise variable was generated from a standard normal distribution (mean=0, standard deviation=1), which is consistent with the scaled data processed by Z-Score Normalization. To understand how feature dilution increases with more noise, we use the Bank Marketing dataset as an example. With 20 noise variables, the ratio of noise added to the 16 original features is more than half; i.e., 20/36 of the features are noise. When the number of noise attributes is increased to 80, about 83% (i.e., 80/96) of the input features are irrelevant, making it challenging for the algorithm to detect signal among noise.

Three key performance measures are used in the experiments, namely the basic response rate, lifted response rate, and lift. The basic response rate is the baseline against which the lifted response rate is compared. The lift, defined as their ratio, quantifies the improvement to which the basic response rate is lifted. The lift is beneficial for evaluating conversion in marketing contexts.

4.4 Performance When $k = 10$

Table 2 shows the results of applying the proposed method on the two datasets, with the lifted response rates significantly higher than the basic response rates (i.e., $r_L > r_B$) and the effective Lift greater than 1 (i.e., $Lift > 1$) for both datasets ($p < 0.05$). The estimates also accurately preserve these relationships (i.e., $\hat{r}_B > \hat{r}_L$, and $\widehat{Lift} > 1$), confirming the method's reliability in effectively lifting the response rates across the two datasets.

We use $k = 10$ because micro-clustering can better handle imbalanced high-dimensional data, enabling effective cluster ranking by external criteria (e.g., response rate). While traditional methods prioritise identifying "true" cluster counts, micro-clustering approach proves more robust for imbalanced distributions, which will be illustrated in the later results.

Table 2: The improved response rates after applying the proposed method. The results presented are mean ± standard deviation computed based on 30 runs of k-means clustering.

Metric	Online shopping	Bank marketing
r_B	0.155	0.117
$\hat{r}_B$	0.151 ± 0.014	0.125 ± 0.009
r_L	0.299 ± 0.018	0.202 ± 0.008
$\hat{r}_L$	0.306 ± 0.025	0.241 ± 0.032
$Lift$	1.932 ± 0.119	1.731 ± 0.071
$\widehat{Lift}$	2.033 ± 0.172	1.932 ± 0.148

Table 3: Details of response estimates derived from one of the clustering runs applied on the Bank Marketing dataset. The estimated basic response rate is $\hat{r}_B = 5548.64/45211 = 0.123$. Clusters 4 and 10 are high-propensity clusters. r_i is only known after an entire cluster i is fully surveyed. In practice, the values of r_i and its derivations are known for the high-propensity clusters only.

Cluster	n_i	$\sum_{j=1}^{n_i'} y_j$	$\hat{r}_i$	$\lfloor \hat{r}_i \cdot n_i \rfloor$	r_i	$r_i n_i$
cluster-1	7898	6	0.060	473	0.035	278
cluster-2	2977	6	0.060	178	0.074	220
cluster-3	1853	14	0.140	259	0.154	286
cluster-4	4350	26	0.260	1131	0.216	940
cluster-5	3294	14	0.140	461	0.100	331
cluster-6	4626	10	0.100	462	0.048	220
cluster-7	5857	5	0.050	292	0.086	506
cluster-8	7242	16	0.160	1158	0.189	1369
cluster-9	3891	10	0.100	389	0.105	410
cluster-10	3223	23	0.230	741	0.226	729
All clusters	45,211	130	0.123	5548	0.117	5289
Clusters 4 & 10	7573	49	0.247	1872	0.220	1669

Bank Marketing Example: One key advantage of applying the proposed method is cost savings in comparison to using traditional mass marketing. For example, Table 2 shows that the actual Lifts are 1.932 and 1.731 for the Online Shopping and Bank Marketing datasets, respectively. This is achieved by engaging the remaining customers (who were not contacted previously during the sampling stage) within the high-propensity clusters.

Table 3 shows the estimated responses of customers within the samples stratified by the 10 clusters, which were generated from one of the clustering runs applied on the Bank Marketing dataset. For example, cluster-4 has 26 positive responses out of the 100 records in the cluster subsample; thus, the estimated cluster response rate is 0.26. This cluster's estimated number of positive responses is $0.26 \times 4350 = 1131$ customers. The estimated basic response rate is obtained by summing up the values under the $\hat{r}_i \cdot n_i$ column, and then dividing by the total data size of 45,211, resulting in 0.123, which is quite close to the actual response rate of 0.117.

Assume that the average product cost per customer is \$100, the telemarketing cost per customer is \$5, and the average revenue per customer for every successful campaign engagement is \$150. The costs and benefits of applying the proposed approach on the Bank Marketing dataset are presented in Table 4.

The first column of Table 4 shows the Initial Pilot Campaign conducted on 1000 customers in samples stratified by 10 clusters, where each cluster has 100 customers. This can be seen as a small-scale mass marketing to obtain the initial estimates of cluster response rates, basic response rates, and high-propensity clusters.

The second column of Table 4 shows the targeted campaign on customers remaining within each high-propensity cluster. The total number of remaining customers engaged in clusters 4 and 10 is 7573 minus 200 customers already engaged in the pilot campaign. Thus, this high-propensity cluster campaign resulted in a profit of \$44,135. By conducting the initial pilot on all

Table 4: Cost and benefit analysis of the initial pilot campaign based on 10 sampled clusters (first column), campaign conducted to remaining customers in high-propensity clusters (second column), and traditional mass marketing (third column).

	Initial pilot campaign	High-propensity clusters	Mass marketing
Data Size	1000	7373	45,211
Telemarketing Cost ($5/pax)	$5,000	$36,865	$226,055
# Positive Respondents	130	1620	5289
Product Cost ($100/pax)	$13,000	$162,000	$528,900
Revenue ($150/pax)	$19,500	$243,000	$793,350
Profit	$1,500	$44,135	$38,395
Return on Investment	7.69%	18.16%	4.84%
Work hours (5 minutes/pax)	83.3	614.4	3767.6

clusters and then the follow-up campaigns on clusters 4 and 10, the resulting combined profit is $45,635 (i.e., $44,135 + $1500).

The third column of Table 4 shows the results of applying traditional mass marketing on the whole dataset. The profit is $38,395, which is lower than that generated by the proposed approach.

Another important consideration is the Return on Investment (ROI), where Table 4 shows that the proposed approach is the most profitable (ROI = 18.16%) relative to its cost. Collectively, the initial pilot and subsequent high-propensity clusters target only 8373 out of 45211 customers, significantly reducing the outreach efforts compared to contacting 45,211 customers in mass marketing.

The high ROI implies efficiency, where resources (e.g., time) are used effectively to generate profit. If it takes on average five minutes to engage one customer, our proposed method will require about 698 work hours. On the other hand, traditional mass marketing requires 3768 work hours. Thus, the proposed approach achieves better outcomes yet consumes substantially less time and resources than traditional mass marketing.

Online Shopping Example: Unlike the traditional bank marketing case, online shopping platforms are more cost-effective and automated in customer engagement, since they can reach many online shoppers electronically. However, massive online dissemination of advertisements can also lead to undesirable outcomes, such as brand fatigue. Excessive attention-seeking advertisements can alienate customers rather than build a genuine connection.

The online shopping dataset contains online users' shopping behaviours collected via Google Analytics (Google LLC, 2025). The profiles of the clusters explain why the proposed approach improves the response rate. Table 5 shows the online shopping profiles of ten clusters ordered by the response rates from top to bottom in descending order. We can gain insight into how clustering differentiates high-propensity clusters (i.e., clusters 2, 9, and 6) from the low-propensity clusters (i.e., clusters 10, 1, and 8).

Table 5 shows that the low-propensity clusters contain sessions with substantially higher bounce rates of 18.6% on average (weighted by respective cluster sizes), compared to the high-propensity clusters of only 0.64%. This means that sessions in low-propensity clusters tend to have very short page visit durations and lack further user interaction after a page is loaded.

Table 5: Cluster-level summary of online shopper behavior, where clusters are grouped into high- and low-propensity segments based on response rate $\mathbf{r_L}$ and engagement indicators.

Cluster	$\mathbf{n_i}$	$\mathbf{r_i n_i}$	$\mathbf{r_L}$	**Bounce**	**Exit**	**PageValue**
cluster-2	58	16	0.276	0.026	0.045	27.627
cluster-9	1765	456	0.258	0.010	0.031	6.650
cluster-6	1665	421	0.253	0.002	0.018	10.937
Top-Propensity	3488	893	0.256	0.64%	2.50%	9.045
cluster-7	2256	393	0.174	0.011	0.032	6.043
cluster-3	1292	165	0.128	0.010	0.032	6.386
cluster-5	2273	253	0.111	0.011	0.036	3.954
cluster-4	2172	201	0.093	0.014	0.039	4.693
cluster-10	180	2	0.011	0.185	0.192	0.000
cluster-1	564	1	0.002	0.185	0.191	0.000
cluster-8	105	0	0.000	0.191	0.195	0.000
Low-Propensity	849	3	0.004	18.57%	19.17%	0.000

A high bounce rate suggests a potential mismatch between page content and the intention of shoppers.

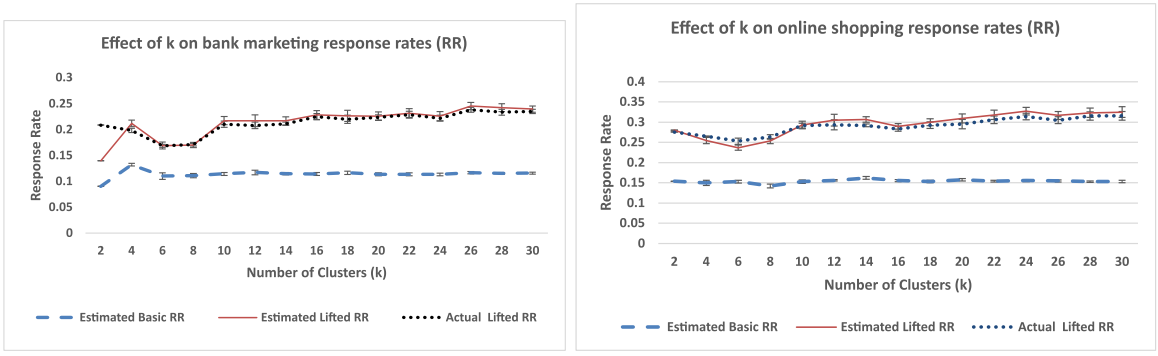
Table 5 also highlights that sessions in low-propensity clusters tend to have a higher exit rate of 19.7%, compared to the high-propensity clusters of only 2.5% on average. A high exit rate suggests that low-propensity shoppers do not find the shopping platform engaging or relevant enough, resulting in them leaving the platform.

Further analysis of Table 5 reveals that sessions in low-propensity clusters have zero PageValue, while the high-propensity clusters have an average PageValue of 9.05. In this dataset, PageValue is the average monetary contribution of a web page that an online user visited, right before completing a transaction on the e-commerce website. A webpage with a high PageValue tends to influence users to complete a transaction.

In summary, the results highlight a big difference in profiles between the high- and low-propensity clusters. The high-propensity clusters offer high conversion potential (25.6% average lifted response rate), robust user engagement (low bounce and exit rates), and high monetization potential (high PageValue). In contrast, the low-propensity clusters exhibit negligible conversions (0.4% average lifted response rate), weak shopper engagement (high bounce and exit rates), and no measurable revenue contribution (zero PageValue). By clustering online shoppers according to their profiles, the proposed approach can minimise unnecessary and ineffective outreach. This is achieved by focusing on engaging high-propensity customer groups, so as to foster stronger customer relationships.

4.5 Effects of Varying Number of Clusters

Figure 2 demonstrates that the lifted response rate tends to stabilise when there are 10 or more clusters. This is expected because generating more clusters creates smaller, more homogeneous micro-clusters, thereby refining response rate differentiation.



(a) Response rates across clusters for Bank Marketing. (b) Response rates across clusters for Online Shopping.

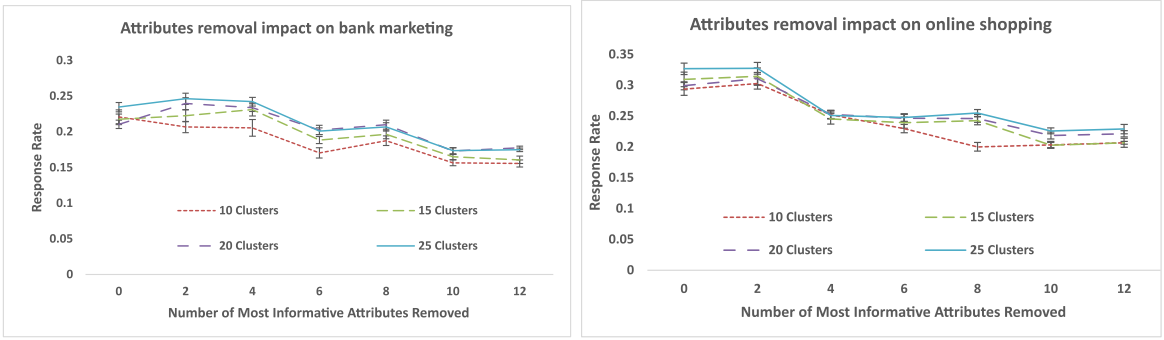
Figure 2: Response rates tend to stabilise and improve with more generated clusters for the Bank Marketing and Online Shopping datasets.

For Bank Marketing, the lifted response rate plateaus at 0.22–0.24 (vs. baseline 0.117; $p < 0.05$), while Online Shopping achieves 0.29–0.33 (vs. baseline 0.155; $p < 0.05$). The small error bars demonstrate tight 95% confidence intervals. This indicates that the lifted response rate is significantly higher than the basic response rate across all cluster configurations.

4.6 Effects of Removing Informative Attributes

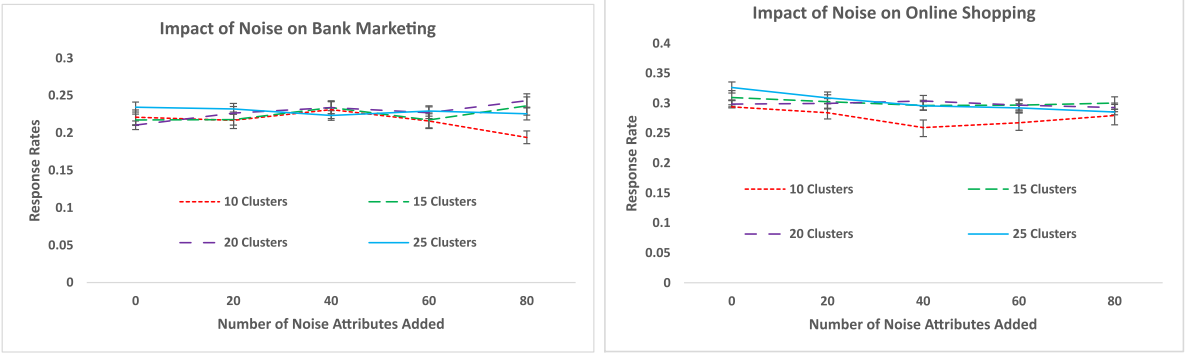
Figure 3 shows that the response rates decline as informative attributes are progressively removed. When 10 clusters are generated, the lifted response rates for Bank Marketing and Online Shopping reduce from 0.22 to 0.16, and 0.30 to 0.20, respectively. This decline is expected because removing informative attributes inevitably loses information and reduces cluster distinctiveness.

Figure 3 demonstrates that the method maintains stable performance across different numbers of clusters, even when stressed with incomplete features. This is evident from the error bars (95% confidence intervals) in Figure 3, which indicate that the worst-case lifted response rates



(a) Bank Marketing response rates for different numbers of removed attributes. (b) Online Shopping response rates for different numbers of removed attributes.

Figure 3: Response rates tend to reduce as more informative attributes are removed from the Bank Marketing and Online Shopping datasets.



(a) Bank Marketing response rates for different numbers of noise attributes.

(b) Online Shopping response rates for different numbers of noise attributes.

Figure 4: Comparison of response rates for different numbers of noise attributes added to Bank Marketing and Online Shopping datasets.

(0.16 and 0.20) remain statistically significant compared to their respective baseline response rates (0.117 and 0.155). In practice, however, their practical significance should be assessed by considering related costs and benefits.

4.7 Effects of Adding Noise Attributes on Lifted Response Rates

Figure 4 demonstrates that including noise attributes leads to a marginal response rate reduction in the Online Shopping problem. This can happen because noise and irrelevant attributes may distort the true distances between data points, thereby affecting clustering quality.

Interestingly, the reduction in lifted response rates is less severe than expected, with most error bars staying well above the baseline response rates. We attribute this robustness of the method to two key factors: First, the datasets may contain naturally well-separated clusters. Second, our method tends to use a relatively large number of micro-clusters. These micro-clusters capture dense and meaningful subsets of the data. In addition, noisy data points tend to average out across many smaller clusters. This effect becomes evident when comparing different cluster counts: solutions with 15 clusters or more tend to have similar performance, while solutions with 10 clusters display greater variability, as seen in Figure 4.

5 Concluding Remarks

This paper proposed a label-efficient, cluster-based active sampling method for response modeling in scenarios where ground-truth labels are scarce or unavailable. By combining stratified sampling with micro-clustering, the proposed method identified high-propensity customers with marginal label acquisition costs. Our results showed that the method lifted the basic response rates by at least 1.7 times for Bank Marketing, and 1.9 times for Online Shopping.

For Bank Marketing, the proposed method generated higher profits with fewer resources compared to traditional mass marketing. For Online Shopping, the strategy of ranking micro-clusters by response rates enabled the semantic interpretation of cluster structures. The results revealed clear distinctions between high- and low-propensity clusters. For example, the stark differences in bounce rates and page values provide actionable insights for digital marketing.

Our analysis uncovered an interesting finding: micro-clustering detected high-propensity subgroups in a dataset with up to 83% noise variables. This unexpected tolerance to noise suggests that there might be an important but poorly understood property in the proposed method. Uncovering the underlying mechanism may benefit analyses in noisy and high-dimensional data space.

One limitation of the proposed method lies in determining the optimal number of micro-clusters (k). Although we empirically recommend $k \geqslant 10$, this setting warrants a more systematic evaluation. One possible evaluation method is to rank the clusters in descending order of their response rates: an appropriate k should usually show noticeable differences between the profiles of high- and low-propensity clusters.

Since the proposed approach is label-efficient, it has the potential for broad applications in different domains. But before this could happen, its adoption needs to be facilitated. Our next step is to develop a practitioner's guide for applying this methodology in different industries. For example, it will be useful to develop a simple statistical rule-of-thumb for practitioners to gauge what makes a high-propensity cluster, without having to apply statistical tests. Another plan is to develop a user-friendly application to automate the analysis.

Supplementary Material

The Python notebook containing the implementation of the proposed method is available at the following link: https://colab.research.google.com/drive/1IG-9N7iakfPUUKnIskKYH2kbs_F6sdgP?usp=sharing.

Additionally, the datasets used in this study, where features are ordered by their importance (from left to right), can be accessed via: https://drive.google.com/drive/folders/1WE8A0aZ-cKLJ45hRDMFH20CczwZ2wiWh?usp=sharing.

Acknowledgement

The author thanks his undergraduate student, Ms. Thiri Cho, for her assistance in early exploratory work during the initial stages of this project. The author is grateful to the editor, associate editor, and referees for their constructive comments which has led to significant improvement of this paper.

References

Ali A, Abd Razak S, Othman SH, Eisa TAE, Al-Dhaqm A, Nasser M, et al. (2022). Financial fraud detection based on machine learning: A systematic literature review. *Applied Sciences*, 12(19): 9637. https://doi.org/10.3390/app12199637

Baesens B (2004). Developing intelligent systems for credit scoring using machine learning techniques. *Ph.D. Thesis, Katholieke Universiteit Leuven, Belgium.*

Chaudhuri N, Gupta G, Vamsi V, Bose I (2021). On the platform but will they buy? Predicting customers' purchase behavior using deep learning. *Decision Support Systems*, 149: 113622. https://doi.org/10.1016/j.dss.2021.113622

Emtiyaz S, Keyvanpour M (2011). Customers behavior modeling by semi-supervised learning in customer relationship management. arXiv preprint: https://arxiv.org/abs/1201.1670.

Gönül FF, Hofstede FT (2006). How to compute optimal catalog mailing decisions. *Marketing Science*, 25(1): 65–74. Published online: January 1, 2006. https://doi.org/10.1287/mksc.1050.0136

Google LLC (2025). Google analytics. Web analytics platform.

Hanssens DM, Leeflang PSH, Wittink DR (2005). Market response models and marketing practice. *UCLA Anderson School of Management*.

Haron NHB (2022). Stratified sampling using cluster analysis. *AIP Conference Proceedings*, 2472(1): 050012.

Haughton D, Oulabi S (1993). Direct marketing modeling with CART and CHAID. *Journal of Direct Marketing*, 7(3): 16–26. 11 pages. https://doi.org/10.1002/dir.4000070305

He H, Garcia EA (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9): 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Housden M, Thomas B (2002). *Direct Marketing in Practice*, 1st edition. Routledge, London. EBook published 27 April 2012.

Kang P, Cho S, MacLachlan DL (2012). Improved response modeling based on clustering, under-sampling, and ensemble. *Expert Systems with Applications*, 39(8): 6738–6753. https://doi.org/10.1016/j.eswa.2011.12.028

Lee HJ, Shin H, Hwang SS, Cho S, MacLachlan D (2010). Semi-supervised response modeling. *Journal of Interactive Marketing*, 24(1): 42–54. https://doi.org/10.1016/j.intmar.2009.10.004

Mohammed Amine Naji S, El Filali S, Aarika K, Benlahmar EH, Ait Abdelouhahid R, Debauche O (2021). Machine learning algorithms for breast cancer prediction and diagnosis. *Procedia Computer Science*, 191: 487–492. https://doi.org/10.1016/j.procs.2021.07.062

Moro S, Rita P, Cortez P (2014). Bank marketing. *UCI Machine Learning Repository*.

Sakar C, Kastro Y (2018). Online shoppers purchasing intention dataset. *UCI Machine Learning Repository*.

Thomas AR (2007). The end of mass marketing: Or, why all successful marketing is now direct marketing. *Direct Marketing: An International Journal*, 1(1): 6–16. https://doi.org/10.1108/17505930710734107

Tipton E (2013). Stratified sampling using cluster analysis: A sample selection strategy for improved generalizations from experiments. *Evaluation Review*, 37(2): 109–139. https://doi.org/10.1177/0193841X13516324

Tékouabou SCK, Gherghina SC, Toulni H, Neves Mata P, Mata MN, Martins JM (2022). A machine learning framework towards bank telemarketing prediction. *Journal of Risk and Financial Management*, 15(6): 269. https://doi.org/10.3390/jrfm15060269

Yan X, Nazmi S, Gebru B, et al. (2022). A clustering-based active learning method to query informative and representative samples. *Applied Intelligence*, 52: 13250–13267. https://doi.org/10.1007/s10489-021-03139-y