International Scientific Journal of Engineering and Management
Volume: 04 Issue: 05 | May – 2025
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM03377

Stratified sampling in Cohort-based Data for Machine Learning Model Development

Vaibhav Tummalapalli
Atlanta, USA
vaibhav.tummalapalli21@gmail.com

***Abstract*—Cohort-based data is a prevalent structure in many industries, enabling longitudinal analyses and tracking customer behaviors over time. However, sampling such data for model development presents unique challenges, especially when events (e.g., responses, purchases) are unevenly distributed across cohorts. Random sampling can introduce biases, leading to models that fail to generalize. This paper presents a stratified sampling framework designed to maintain the proportional representation of events and non-events within each cohort, even when oversampling or undersampling is applied. The approach ensures stable and unbiased models, offering insights into practical implementation and evaluation metrics**

Keywords—Random, Stratified Sampling, Machine Learning, Cohort analysis, Class Imbalance, Sampling Bias.

I. INTRODUCTION

Cohort-based data structures are widely adopted in domains such as finance, marketing, and healthcare to capture and analyze temporal trends in customer behavior, product usage, or treatment outcomes. By grouping observations based on a shared characteristic (e.g., acquisition date, diagnosis month, or campaign exposure), cohorts enable time-sensitive and context-aware insights. However, while this structure supports more meaningful segmentation and trend analysis, it introduces unique challenges during machine learning model development—particularly when the target variable (event) rates vary significantly across cohorts.

One critical modeling challenge arises from imbalanced class distributions that may not be uniform across cohorts. For instance, in a churn prediction model across monthly acquisition cohorts, newer cohorts may have lower churn rates due to insufficient observation periods, whereas older cohorts may show higher rates simply due to extended follow-up. If not properly handled, this variation can lead to biased learning, overfitting, and reduced generalizability.

In traditional practice, random sampling is often employed to select training and validation datasets. However, this approach can disrupt the inherent event/non-event distribution within each cohort, effectively diluting cohort-specific dynamics. This can mislead the model into learning patterns that do not generalize well across time or customer segments.

To address this, we propose a stratified sampling framework specifically designed for cohort-based datasets. This framework ensures that sampling occurs within each cohort stratum, preserving the relative proportions of event and non-event cases. By respecting the distributional properties of each cohort, stratified sampling improves:

- Model calibration, by aligning training data more closely with production distributions;
- Stability across time, reducing temporal bias in evaluation metrics;
- Fairness and interpretability, ensuring no cohort is under- or over-represented in model training.

Unlike simple stratified sampling across the whole dataset, our approach applies stratification hierarchically: first by cohort, then by class label. This two-level stratification is particularly effective in longitudinal or delayed-response scenarios where the outcome of interest (e.g., conversion, repurchase, readmission) evolves with time.

This paper outlines the theoretical foundation, practical implementation, and empirical benefits of this approach. We demonstrate, using real-world marketing and healthcare datasets, that cohort-respecting stratified sampling leads to more accurate, robust, and interpretable models, particularly in the presence of class imbalance and shifting cohort characteristics over time [1], [2].

II. COHORT SET UP

A. Cohort Definition

Cohorts refer to data groupings based on shared characteristics or temporal boundaries. For instance In a time-based cohort design, customers are segmented according to their entry or engagement period—commonly by month or quarter. For instance, customers acquired in January 2023 would belong to the "Jan 2023" cohort, those acquired in February to the "Feb 2023" cohort, and so on. This temporal anchoring allows for both comparative performance tracking and behavioral modeling across similar lifecycle stages.

Each cohort consists of two key components:

Observation Window: This period precedes the cohort start date and is used to aggregate historical data to derive features. These may include:

- **Recency:** Time since last interaction or transaction
- **Frequency:** Count of interactions or purchases
- **Monetary Value:** Amount spent, average transaction value, etc.

International Scientific Journal of Engineering and Management
Volume: 04 Issue: 05 | May – 2025
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM03377

Performance Window: The time following the cohort date during which the **event of interest** is tracked. This may involve a binary outcome such as purchase, campaign response, churn, or clinical outcome (e.g., readmission in healthcare).

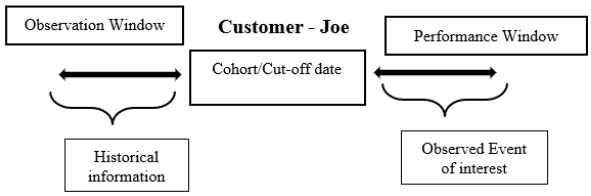

Fig 1. Cohort Framework

Figure 1 presents a visual depiction of how cohort-based evaluation operates over time for an individual customer. In this case, the customer "Joe" is included in multiple cohorts, each corresponding to a different reference date (e.g., each month). At each cohort time point:

- A fixed observation window looks back to summarize past behavior.
- A forward-looking performance window determines whether the event of interest occurred in the defined future period.

This method is applied systematically to all customers in the population, enabling the model to learn patterns that are consistent across different time points while capturing behavioral dynamics and seasonality.

Benefits of Separating Observation and Performance Windows:

- **Temporal Integrity:** Ensures that only past information is used to predict future outcomes, preventing **data leakage** and enabling realistic simulation of model performance in production.
- **Dynamic Feature Recalculation:** By recalculating features at each cohort point, the model accounts for how customer behavior evolves, improving generalizability across time.
- **Behavioral Trend Analysis:** Enables trend comparisons across cohorts, helping identify how customer response patterns shift due to seasonality, lifecycle, or campaign exposure.

B. Example

Consider a marketing dataset where customers are grouped into quarterly cohorts. Each cohort tracks historical communications, transactions, and demographics (observation window) and measures whether a customer responded to a campaign in the following quarter (performance window).

Table 1 – Cohort wise Events & Non – Events (Example set)

Cohort	Total Customers	Events (Responses)	Non-Events (No Responses)	Event Rate (%)
Q1	10,000	1,000	9,000	10%
Q2	12,000	800	11,200	6.70%
Q3	8,000	1,200	6,800	15%

This setup highlights varying event rates across cohorts, which must be preserved during sampling

III. PROPOSED SAMPLING FRAMEWORK

A. Random Sampling

Random sampling, while simple and widely used, often overlooks the underlying structure and temporal dynamics embedded in cohort-based datasets. In scenarios where data is segmented into cohorts—such as quarterly customer acquisition groups—this approach can lead to unintended sampling bias, particularly when the event rates vary significantly across cohorts.

Illustrative Example: Consider a dataset where customers are grouped into quarterly cohorts, and Q2 exhibits a campaign response rate of only 6.7%. If random sampling disproportionately selects event cases (responders) from Q2 to correct for class imbalance at the global level, it can artificially inflate the event proportion within that cohort. This misrepresentation can mislead the model into overestimating the likelihood of future events in similar Q2-like conditions, resulting in overfitting.

Consequences of Ignoring Cohort-Level Distributions:

- **Bias in Model Training:** Random sampling can distort the true event-to-non-event ratio within each cohort, especially in rare event scenarios. This results in the model learning from a distribution that does not reflect real-world conditions, degrading its predictive power.
- **Model Instability Across Time:** A model trained on a dataset that fails to respect cohort structure may exhibit high variance in performance when evaluated across different time periods or customer segments. This is particularly problematic in marketing or finance, where seasonal trends, economic cycles, **or** customer lifecycle stages heavily influence outcomes.
- **Loss of Cohort Integrity:** Temporal and behavioral signals unique to each cohort—such as changes in customer engagement, channel preference, or demographic shifts—can be diluted or erased when cohort identities are ignored. This undermines the ability to extract actionable insights and makes downstream campaign optimization less effective.

International Scientific Journal of Engineering and Management
Volume: 04 Issue: 05 | May – 2025
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM03377

Need for Cohort-Aware Sampling

- To mitigate these issues, it is essential to adopt stratified sampling techniques that maintain the original event/non-event ratios within each cohort. This preserves the integrity of the cohort structure and ensures that models are trained on representative and temporally aligned samples, leading to:
    1. Better generalization to future cohorts,
    2. More reliable calibration of predicted probabilities,
    3. Improved interpretability and business alignment.
- **Loss of Cohort Integrity**: Ignoring cohort structures can obscure temporal trends or demographic differences

B. Stratified Sampling

Stratified sampling provides a principled framework to preserve the statistical structure of data during the sampling process, particularly when the dataset is segmented into meaningful subgroups or strata, such as temporal cohorts. In cohort-based machine learning applications, this method ensures that both event and non-event proportions are maintained within each cohort, regardless of whether oversampling or undersampling techniques are used to address class imbalance.

This approach is grounded in seminal works by Cochran (1977) and Rao (1965), who emphasized the importance of proportional representation across strata to minimize bias and variance in survey and experimental data [3], [4]. When applied to machine learning, the same logic holds: maintaining cohort-level distributions supports model stability, calibration, and generalization—especially when working with imbalanced classification problems where event rates differ across cohorts.

Below are the steps to sample data in a cohort set up.

- **Compute Cohort-Level Proportions**: For each cohort, calculate the event and non-event rate:

$$\text{Event Rate} = \frac{\text{No. of events in the cohort}}{\text{Total Customers in the Cohort}}$$

- **Apply Sampling**:
    - **Oversample Events**: Within each cohort, replicate or synthetically generate additional event (positive class) observations to increase their representation. This is often necessary in cohorts with extremely low event rates (e.g., less than 5%) to ensure sufficient signal for the learning algorithm.
    - **Under sample Non-Events**: In cohorts where non-event observations vastly outnumber events, randomly down sample non-events while keeping the event-to-non-event ratio consistent with the original cohort distribution or adjusting it slightly for learning objectives (such as a 1:1 ratio for balanced training).

Advantages of Cohort-Aware Stratified Sampling

- **Preservation of temporal and behavioral patterns:** This approach maintains the integrity of time-based or behaviorally distinct cohorts, which is critical for understanding lifecycle effects and evaluating seasonal or campaign-driven changes.
- **Improved model calibration**: Since sampling aligns with real-world distributions, predicted probabilities are more likely to reflect true risk or response likelihood across cohorts.
- **Enhanced generalizability**: Models trained using cohort-respecting samples are less prone to overfitting and more robust when deployed on future or unseen cohorts.

C. Example

Table 2 – Conversion/Event Rate by Cohort

Cohort	Non-Events	Events	Total	Conversion
1-Jan-18	1,104,008	562,515	1,666,523	34%
1-Apr-18	1,161,012	567,563	1,728,575	33%
1-Jul-18	1,217,355	569,421	1,786,776	32%
1-Oct-18	1,280,360	573,441	1,853,801	31%
1-Jan-19	1,343,081	570,415	1,913,496	30%
1-Apr-19	1,401,135	567,004	1,968,139	29%
1-Jul-19	1,463,713	560,881	2,024,594	28%
Total	8,970,664	3,971,240	12,940,000	31%

Assume a dataset with an overall event rate of 31%, representing a total population of 12 million customers. As shown in the figure above, this population includes approximately 8.9 million non-events and 3.9 million events, distributed across multiple cohorts ranging from January 2018 to July 2019.

In many real-world scenarios, training a machine learning model on the entire population may be computationally infeasible due to processing time, memory constraints, or operational limits. Therefore, it is common practice to select a representative subset of the data that reflects the characteristics of the full population. However, to ensure statistical fidelity and preserve cohort-level behavioral patterns, the sampling must maintain both the global class proportions and the cohort-wise structure.

For example, consider the goal of extracting a sample consisting of 300,000 event cases and 400,000 non-event cases. To ensure proportional representation across all cohorts, especially when each cohort has varying sizes and event rates, the sampling strategy should follow a structured process as outlined below:

International Scientific Journal of Engineering and Management
Volume: 04 Issue: 05 | May – 2025
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata
ISSN: 2583-6129
DOI: 10.55041/ISJEM03377

Calculate Cohort-Level Proportions:

Proportion of Events in Cohort i:

$$P_{\text{events, cohort } i} = \frac{\text{Number of Events in Cohort } i}{\text{Total Events in the Population}}$$

Proportion of Non-Events in Cohort i:

$$P_{\text{events, cohort } i} = \frac{\text{Number of Events in Cohort } i}{\text{Total Events in the Population}}$$

Determine Sample Sizes for Each Cohort:

Event Sample for Cohort i = $P_{\text{events, cohort i}} \times 300{,}000$

Non-Event Sample for Cohort i = $P_{\text{non-events, cohort i}} \times 400{,}000$

Sample from Each Cohort: Once the sample counts are determined for each cohort, the required number of events and non-events is randomly selected within each cohort to create a balanced and representative dataset for model development. This ensures that the cohort-level distributions are preserved, addressing any potential biases introduced by imbalanced or random sampling approaches.

Table 3 – Samples by Cohort

Cohort	Proportion of events	Proportion of non-events	Event Sample	Non-Event Sample	Total
1-Jan-18	12%	14%	36,921	56,659	93,579
1-Apr-18	13%	14%	38,827	57,167	95,994
1-Jul-18	14%	14%	40,711	57,354	98,066
1-Oct-18	14%	14%	42,818	57,759	100,578
1-Jan-19	15%	14%	44,916	57,455	102,370
1-Apr-19	16%	14%	46,857	57,111	103,968
1-Jul-19	16%	14%	48,950	56,494	105,444
Total	100%	100%	300,000	400,000	700,000

By following the outlined sampling steps, we calculate the cohort-specific sample counts presented in the table above.

D. Evaluation

After implementing cohort-aware stratified sampling, it is essential to evaluate both the **sampling fidelity** and the **modeling impact**. The evaluation process focuses on validating the representativeness of the sampled data and measuring improvements in predictive performance and generalizability. The following components outline a robust evaluation framework:

Cohort Proportionality

Begin by validating that the sampled dataset maintains the original event/non-event ratios within each cohort. This step confirms that the stratified sampling process correctly preserved the distributional characteristics of the full population. Specifically:

- For each cohort, compute the sampled event rate and compare it to the original cohort event rate.
- Calculate aggregate deviation metrics (e.g., mean absolute error across cohorts) to quantify sampling accuracy.
- Visualize the event rate per cohort before and after sampling using bar charts or line plots to visually inspect alignment.

This ensures the structural integrity of cohort patterns is not compromised, which is critical for models trained on time-segmented data.

Model Performance

Next, assess whether stratified sampling leads to improvements in model performance compared to random sampling. Train two versions of the model—one on the cohort-stratified sample and the other on a randomly drawn sample—and compare their outputs using standard classification metrics:

- **ROC-AUC** (Receiver Operating Characteristic - Area Under Curve): Measures overall discriminative ability.
- **Lift at k%**: Evaluates how well the model ranks true positives in the top-scored decile(s), which is especially important in marketing and rare event scenarios.
- **Precision and Recall**: To examine accuracy and coverage on the positive class.

Stratified sampling is expected to improve model calibration, reduce overfitting, and yield more stable predictions, particularly in cohorts with low event rates.

Back Testing on Cohort-Holdout Set

To further assess generalizability, conduct **back testing** using a **holdout dataset** that retains the original cohort structures and has not been used during sampling or training. This simulates real-world deployment and helps answer critical questions:

- Does the model generalize well to future cohorts?
- Are temporal patterns (e.g., shifts in behavior, campaign effects) correctly captured?
- Is the model robust to varying event rates across time?

Back testing should replicate production scenarios, such as predicting outcomes for a future quarter based on prior behavior, allowing a realistic evaluation of cohort-aware model performance.

IV. CONCLUSION

This paper introduces a stratified sampling framework for cohort-based data, ensuring that cohort-level event distributions are preserved. By maintaining proportionality, the approach addresses imbalances and improves model stability, making it particularly suited for applications in marketing, finance, and other domains reliant on longitudinal data. Future work could explore integrating this framework with advanced synthetic sampling techniques and domain-specific weighting strategies.

ISSN: 2583-6129
DOI: 10.55041/ISJEM03377

REFERENCES

[1] F. J. Fowler, *Survey Research Methods*, 5th ed., SAGE, 2014.
[2] P. S. Levy and S. Lemeshow, *Sampling of Populations: Methods and Applications*, 4th ed., Wiley, 2008.
[3] C. R. Rao, *Linear Statistical Inference and Its Applications*, 2nd ed., Wiley, 2002.
[4] W. G. Cochran, *Sampling Techniques*, 3rd ed., Wiley, 1977.