INTERNATIONAL JOURNAL OF INTELLIGENT DATA AND MACHINE LEARNING (IJIDML)

eISSN: 3087-4262

Volume. 03, Issue. 03, pp. 05-17, March 2026

Cohort-Based Segmentation Framework for Machine Learning: Structuring Temporal Data for Enhanced Feature Engineering

Vaibhav Tummalapalli
Independent Researcher, Atlanta, USA

Article received: 05/01/2026, Article Accepted: 08/02/2026, Article Published: 13/03/2026
DOI: - https://doi.org/10.55640/ijidml-v03i03-02

ABSTRACT

Cohort-based segmentation is a well-established method for structuring customer data around time-based reference points, enabling causal inference and temporal feature engineering in marketing analytics. While extensively applied in subscription and retail loyalty contexts, its use in transactional service environments such as automotive aftersales remains underexplored. This paper addresses this gap by proposing a structured cohort framework tailored to irregular, discretionary service interactions, defining clear observation and outcome windows to enable robust engineering of recency, frequency, and monetary (RFM) features while avoiding data leakage. A real-world case study demonstrates the framework's practical value, achieving a lift of 2.7 in the top decile and consistent capture rates across cohorts. These results highlight the approach's ability to improve targeting precision, uncover temporal trends (including COVID-19 disruptions), and support marketing strategies for customer retention and engagement in industries with low-frequency, high-value transactions

KEYWORDS

Predictive/Propensity Modeling, Machine Learning, Cohort Analysis, Trend Analysis, Seasonality Detection, Temporal Feature Engineering, Customer Segmentation, Marketing Analytics, and Automotive Marketing.

1. INTRODUCTION

Predictive modeling in marketing analytics relies on well-structured data that accurately reflects customer behavior over time. Temporal structuring is particularly critical for understanding patterns of purchase, usage, or engagement that evolve across the customer lifecycle. One widely used method to achieve this is cohort-based segmentation, which organizes customer data around time-based reference points to enable consistent, causal, and interpretable analysis (Fader & Hardie, 2007; Blattberg et al., 2008).

Cohort analysis has a rich history in marketing analytics. In subscription services, it tracks churn risk, estimates customer lifetime value (CLV), and informs retention strategies by segmenting users based on sign-up dates or campaign cohorts (Gupta & Zeithaml, 2006; Fader & Hardie, 2007). Telecommunications and streaming services use cohort frameworks to analyze how retention patterns respond to marketing interventions (Lemmens & Croux, 2006). Retail loyalty programs evaluate acquisition campaigns and onboarding by segmenting customers based on acquisition dates or promotions (Blattberg et al., 2008; Neslin et al., 2006). The RFM (recency, frequency, monetary) modeling tradition is deeply rooted in these cohort-based approaches (Chen et al., 2009).

Cohort analysis also supports online and app engagement analytics, enabling teams to track retention curves, feature adoption, and monetization over time (Kohavi et al., 2014; Berman, 2018). These applications demonstrate the versatility of cohort frameworks in

capturing dynamic customer behavior and aligning marketing strategies with real-world cycles.

A key benefit of cohort-based segmentation is enforcing temporal consistency in feature engineering. By defining clear observation and outcome windows relative to cohort dates, it prevents data leakage—a common pitfall where future information inadvertently informs model training (Brownlee, 2020). This structuring supports robust extraction of temporal features like recency, frequency, and monetary value, repeatedly validated as strong predictors of purchase and retention propensity (Gupta & Zeithaml, 2006; Chen et al., 2009).

However, much of the existing literature focuses on context with frequent, regular customer interactions. Subscription churn models, loyalty programs, and app engagement analytics benefit from high-frequency event data naturally suited to cohort tracking. In contrast, transactional service environments such as automotive aftersales present unique challenges that have received limited scholarly attention. Service transactions in this domain occur at irregular intervals, involve high-value discretionary spending, and require careful definition of observation windows to engineer meaningful temporal features despite sparse or noisy service histories.

This paper addresses this gap by proposing a structured cohort-based segmentation framework tailored to automotive aftersales marketing. Our approach defines consistent observation and performance windows to enable robust temporal feature engineering while avoiding leakage. By systematically leveraging first-party transactional service data, it supports the development of predictive models for customer return propensity and improves targeting for marketing interventions. A real-world case study demonstrates its effectiveness, achieving a lift of 2.7 in the top decile and consistent capture rates across cohorts. While focused on automotive service marketing, this framework is broadly applicable to other domains with irregular, discretionary customer interactions, such as healthcare scheduling, equipment maintenance, and high-value B2B services.

This framework aligns with established marketing theory around customer lifecycle management, segmentation, and targeted communications by enabling dynamic RFM feature engineering and supporting direct marketing principles of timing and personalization

2. COHORT FRAMEWORK & DATA STRUCTURING

This section describes how cohort-based segmentation organizes transactional data into meaningful, business-aligned time windows for predictive modeling. It explains the design of observation and performance windows while highlighting segmentation levels that can be tailored to specific marketing objectives. This approach aligns with best practices in predictive modeling by enforcing temporal consistency, avoiding data leakage, and supporting interpretable feature engineering (Brownlee, 2020; Fader & Hardie, 2009).

2.1 Defining Observation & Performance Window

Cohort dates help structure data into time-aligned windows that support causal analysis:

- **Observation Window**: Includes all historical data available up to the cohort date. This window is used to extract features such as demographics, transaction history, and temporal behaviors (e.g., recency, frequency, monetary value). By anchoring features to this period, businesses can reliably leverage their first-party transactional data to capture real customer behavior without introducing lookahead bias.
- **Performance Window**: Tracks whether the event of interest (e.g., vehicle purchase or service visit) occurred after the cohort date within a specified time frame. This ensures models are trained with clearly defined temporal cause-and-effect relationships, improving interpretability and predictive validity.

This division simplifies the modeling process, making it easier to assess the impact of historical behavior on future outcomes.

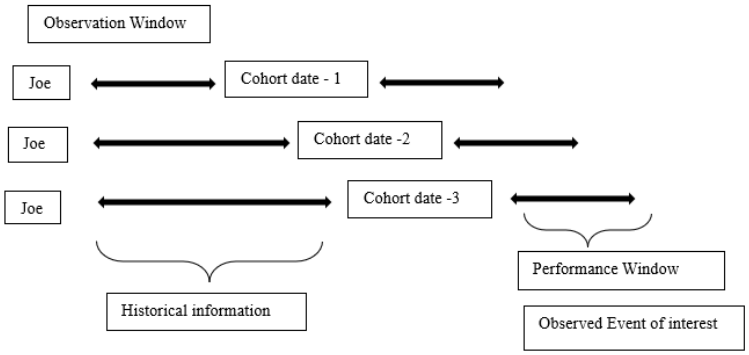

Fig 1. Cohort-Based Framework

Fig 1 provides a visual representation of how cohort-based analysis evaluates customer behavior at various time points. Here's a breakdown of the concept:

- **Multiple Time Points for Analysis**:
  - Here the customer "Joe," is observed at several specific time points (e.g., Cohort date -1, Cohort date -2, Cohort date -3) and the event of interest is assessed in the performance window. This is done for every customer in the population.
  - This approach captures temporal changes in customer behavior, providing insights into patterns and trends over time.

2.2 Customizing Segmentation Levels

Cohort-based segmentation can be tailored to various levels depending on the business objective:

- **VIN Level**: If the goal is to run a marketing campaign targeting specific vehicles, segmentation and modeling are performed at the VIN level. Temporal features such as recency, frequency, and monetary value are calculated for each VIN to represent the vehicle's activity accurately.
- **Customer ID Level**: For broader customer-focused campaigns, segmentation is conducted at the customer ID level. Features are aggregated to reflect a customer's overall behavior across all associated vehicles.
- **Household Level**: When targeting multiple customers within the same household, segmentation aggregates data to reflect household-level patterns. This level of granularity supports campaigns aimed at families or shared ownership scenarios.

These flexible segmentation levels ensure that cohort-based analytics align closely with business goals and campaign objectives.

3. BENEFITS OF COHORT ANALYSIS

Cohort-based segmentation provides a structured framework for organizing customer transactional data over time, supporting predictive modeling and marketing decision-making. While widely applied in subscription and retail loyalty contexts, its extension to irregular, discretionary services such as automotive aftersales addresses a critical gap in marketing analytics by enabling robust temporal feature engineering, trend analysis, and targeted interventions even in low-frequency environments. These benefits align with established marketing theory around customer lifecycle management and segmentation (Blattberg et al., 2008) and are demonstrated in our automotive case study (Section IV).

3.1 Temporal Feature Engineering

Cohort dates allow precise definition of observation windows, ensuring features reflect only historical customer behavior and avoid lookahead bias (Brownlee, 2020). This supports robust engineering of well-established RFM (recency, frequency, monetary) metrics (Fader & Hardie, 2009; Chen et al., 2009):

- **Recency**: Time since last customer action (e.g., service visit, purchase), indicating current engagement levels.
- **Frequency**: Number of events in a defined period, capturing loyalty or attrition risk.

- **Monetary**: Total spending during the observation window, enabling customer value segmentation.

Together, these dynamic features allow predictive models to reflect evolving customer behavior and inform differentiated marketing strategies.

3.2 Trend Analysis & Generalization

Using multiple cohort dates enables businesses to track how customer event rates (e.g., purchases, service visits) evolve over time, accounting for economic conditions, lifecycle stages, or market disruptions. This analysis supports:

- **Trend Detection:** Identifying increasing or decreasing engagement patterns.
- **Proactive Interventions:** Responding to observed declines with targeted incentives or messaging.

Such analysis improves model generalizability by embedding temporal awareness and supports campaign planning that adapts to changing customer behavior

3.3 Seasonality Detection

- **Seasonal Variations:** Cohort trends can reveal seasonality in customer behavior, such as increased vehicle purchases at the end of the year or higher service visits before long weekends.
- **Incorporating Seasonality in Models:** By including cohort-based seasonal insights, predictive models become more aligned with real-world behavior. For instance, a service attrition model can account for peak service times in specific months.

3.4 Segmentation & Personalization

Cohort-based segmentation uncovers differences in customer behavior across groups, supporting targeted marketing and personalized engagement strategies (Blattberg et al., 2008).

- **Cohort-Specific Insights**: Different cohorts exhibit unique behaviors based on factors like timing, promotions, or market conditions. For example, a cohort of customers who purchased vehicles during a promotional period may exhibit higher service loyalty but lower spending per visit compared to a cohort of non-promotional buyers who demonstrate consistent but less frequent engagement.

Examples of Segmentation:

- **Promotional Buyers**: Likely to respond to discounts and incentives. Personalization Strategy: Offer recurring promotions to maintain engagement.
- **Non-Promotional Buyers**: Tend to prioritize quality over discounts. Personalization Strategy: Highlight premium services or long-term benefits.
- **Seasonal Buyers**: Exhibit spikes in purchases or services during specific seasons (e.g., year-end). Personalization Strategy: Target these customers with seasonal campaigns and reminders.

3.5 External Influences

Cohort analysis also enables modeling of macroeconomic and contextual factors:

- **Economic downturns:** Customers may delay purchases or shift to lower-cost services.
- **Product launches:** Drive purchase surges in specific cohorts.
- **Regulatory changes:** Alter service demand (e.g., emissions standards).
- **Market disruptions:** Events like pandemics reshape behavior, visible in response-rate dips.

Incorporating such signals improves model realism and supports risk management

3.6 Model performance Enhancements

Finally, cohort-based segmentation improves predictive modeling performance:

- Reducing Overfitting: Training across multiple cohorts teaches models to generalize better across time, avoiding reliance on idiosyncratic patterns.
- Temporal Validation: Testing on future, held-out cohorts ensure models remain robust to changing conditions, validating their real-world applicability.

These benefits collectively illustrate why cohort-based segmentation remains a foundational method in marketing analytics for understanding customer behavior over time. Our case study further demonstrates these

advantages in the context of automotive after-sales marketing

4. CASE STUDY

The following case study illustrates the application of a cohort-based segmentation framework in an automotive aftersales marketing context to improve predictive modeling and campaign design. This project demonstrates several core benefits of cohort analysis in practice:

- **Temporal Feature Engineering**: Using observation windows to construct recency, frequency, and monetary (RFM) metrics aligned to real customer behavior.
- **Avoidance of Data Leakage**: Ensuring all features were constructed only from data available up to the cohort date.
- **Trend Analysis**: Tracking customer response rates across multiple cohorts to identify temporal patterns and incorporating an explicit COVID-19 flag to account for macroeconomic shocks and market disruptions
- **Model Performance Enhancement**: Validating model generalization over time through temporal back-testing on held-out cohorts.

By highlighting these aspects, the case study provides a practical demonstration of how cohort-based segmentation can deliver structured, reliable, and interpretable predictive models in a real-world marketing context.

4.1 Business Context

This study was conducted for a major automotive client with the objective of identifying retained customers likely to return for service within 60 days of a campaign date. The goal was to develop a predictive model leveraging a cohort-based segmentation framework and machine learning classification to produce stable, interpretable propensity scores (ranging from 0 to 1), indicating each customer's likelihood of returning. By structuring the data around defined cohort dates, the approach ensured temporal consistency and avoided data leakage, supporting established marketing principles of timing, targeting, and personalization. The model was designed to leverage rich historical service transaction data, enabling the business to focus marketing resources on top-scoring, high-propensity customers. This target strategy aims to improve service retention rates, optimize campaign efficiency, and enhance overall customer lifetime value

4.2 Data Characteristics

The modeling dataset was derived from dealership service transaction records and structured using a cohort-based segmentation framework to ensure temporal consistency and avoid data leakage. The eligible population was defined as all customers who had serviced their vehicle at least once in the 12 months prior to each selected cohort date. This approach shows *active customers* are likely to be engaged with service channels.

Cohort Design:

- Ten cohorts were defined, covering the period from November 2018 to February 2021.
- For each cohort date, customers were included if they had at least one service visit in the prior 12 months.
- Data was modeled at the Customer ID + VIN level, enabling predictions at a granularity that supports targeted outreach both to individual customers and specific vehicles. This ensures predictions account for both customer and vehicle-specific behavior.

Sample Size and Response Definition:

- The final modeling dataset comprised approximately 15 million unique observations (Contact ID + VIN + Cohort Date combinations) across the ten cohorts.
- The response variable was defined as whether the customer returned for service within 60 days of the cohort date, with Customer Pay > 0 indicating a positive response
- The overall response rate observed was 21%.

Source Data Source and Feature Aggregation:

- The underlying transaction file included over 200 million individual service records.
- Features were engineered exclusively from data available in the observation window to maintain temporal integrity and avoid leakage.
- Key variables included repair order dates, service types performed, customer payment amounts (labor, parts, miscellaneous), warranty flags, and derived over 200 behavioral features such as recency, frequency, and monetary value.
- An additional COVID-19 flag was included to capture periods of significant market disruption and to account for macroeconomic effects on service behavior.

Response Rate Trends:

- *Figure 2* below illustrates how response rates varied across cohorts.
- A clear dip is observed in **2020**, corresponding to COVID-related lockdowns and reduced customer mobility.

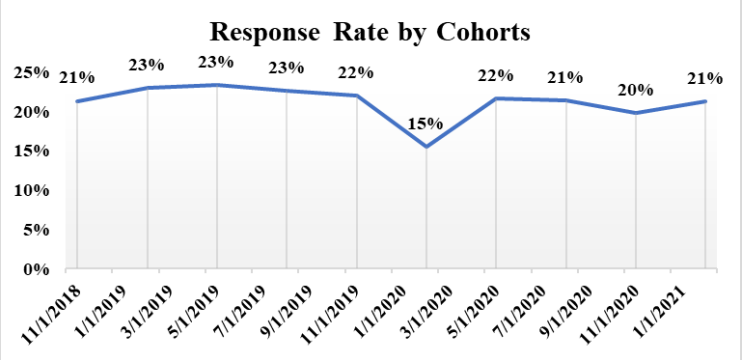

Fig 2. Response Rate by Cohorts

4.3 Modeling Framework

Dependent/Target Variables

The dependent variable represents the outcomes the model seeks to predict for two distinct service types:

- **Outcome 1 (1):** A customer has a repair order with Customer Pay (CP) > 0 within 60 days of the cohort date. This outcome indicates that the customer returned for any general service repair, where the service was paid for out-of-pocket rather than covered by warranty or other programs.
- **Outcome 2 (0):** A customer does not have a repair order within 60 days of the cohort date. This indicates that the customer did not engage with the service center during the specified time frame.
  This variable ensures that the model differentiates between customers who are likely to return for service and those who are not, enabling targeted marketing strategies.

Cohorts

- **Total Cohorts:** The dataset spans 10 cohorts, covering the period from November 2018 to February 2021.
- **Cohorts Used for Modeling:** All 10 cohorts from November 2018 to February 2021 were included in the model training process to capture diverse customer behaviors across time.
- **Back-Test Cohort:** The cohort from May 2021 was reserved exclusively for testing the model's performance on unseen data, ensuring the model's ability to generalize beyond the training dataset.

This design mimics real-world campaign cycles and respects temporal separation between training and validation data, making it broadly applicable to other marketing and service use cases with time-based dynamics

Modeling Unit:

- The modeling unit combines *Customer ID* and *VIN (Vehicle Identification Number)*, creating a granular structure that reflects customer behavior at both individual and vehicle levels.
- This allows for detailed insights into customer-vehicle interactions, enabling personalized service recommendations and campaigns.
- More generally, similar approaches can be applied in other industries by defining appropriate customer or product identifiers (e.g., Account ID, Product SKU) to capture meaningful behavioral patterns at the desired level of granularity.

Model Development Approach:

- A supervised classification model was developed to predict customer propensity within the cohort-based framework. Features were engineered using only data from the observation window, ensuring temporal consistency and avoiding leakage.

- Basic preprocessing steps included stratified sampling to reduce data volume for efficient computation and to maintain event/non-event proportions across training and validation splits, imputing missing values with medians for numeric fields and adding missing indicators for categorical columns, handling outliers through winsorization, and encoding categorical features using one-hot or ordinal encoding based on cardinality and model requirements.
- Feature selection was performed using Information Value, Analysis of Variance (ANOVA), and Chi-Square tests for categorical variables, as well as by leveraging feature importances from machine learning algorithms such as random forests and decision trees.
- Finally, a gradient boosting classifier was chosen for its ability to model complex relationships. Hyperparameter tuning for the gradient boosting model was conducted using grid search with validation on the training data, optimizing for AUC and lift in the top deciles to balance predictive accuracy with marketing prioritization needs. Model performance was evaluated on a held-out back-test cohort to assess generalization to future data.

4.4 Model Metrics

To assess the predictive performance of the model, two key metrics were employed:

Decile Analysis: Predicted propensity scores were ranked and divided into deciles - ten equal-sized groups, each representing 10% of the ranked customer list. The top decile includes the highest 10% of predicted scores, representing customers most likely to respond. This segmentation enables marketers to evaluate model effectiveness in targeting the highest-propensity segments

Lift:

- Lift measures how much better the model is at identifying high-propensity customers compared to random selection.
- For example, a lift of 3 in the top 10% of predictions means that the model identifies three times more likely customers within that segment than a random guess would.
- Higher lift values indicate a more effective model, enabling businesses to focus their resources on the most promising customers.

Capture Rate:

- Capture rate measures the proportion of actual positive outcomes (e.g., customers who returned for service) successfully identified by the model within its top-ranked segments.
- For instance, if the top 20% of predictions include 50% of the actual outcomes, the capture rate for that segment is 50%.
- This metric demonstrates the model's ability to prioritize actionable customer segments and maximize the return on marketing investment.

4.5 Results & Insights

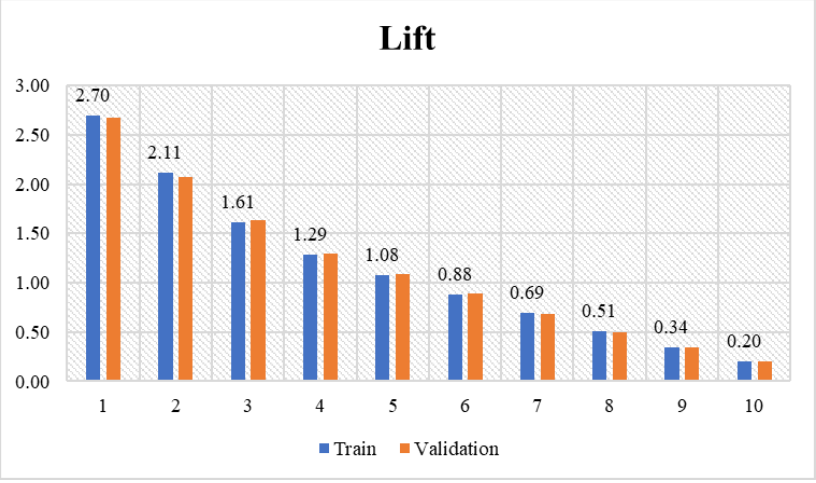

Fig 3. Model Lift: Train & Validation

Lift Analysis:

- Top Decile: Achieves a lift of 2.7, capturing 24% of responders, indicating strong predictive accuracy.
- Top 3 Deciles: Capture 56% of total responders, demonstrating effective prioritization.
- Model Stability: Consistent lift values between training and validation datasets confirm reliability and generalizability.

Overall Response Rate: 24%

Table 1 – Train Results

Training				
Decile	**Non - Responses**	**Responses**	**Total**	**Response Rate**
1	18,190	23,802	41,992	57%
2	23,364	18,628	41,992	44%
3	27,756	14,236	41,992	34%
4	30,622	11,370	41,992	27%
5	32,476	9,516	41,992	23%
6	34,260	7,732	41,992	18%
7	35,907	6,085	41,992	14%
8	37,536	4,456	41,992	11%
9	38,991	3,001	41,992	7%
10	40,241	1,751	41,992	4%
Total	319,342	100,578	419,920	24%

Table 2 – Validation Results

Validation				
Decile	**Non - Responses**	**Responses**	**Total**	**Response Rate**
1	12,274	15,722	27,996	56%
2	15,791	12,204	27,995	44%
3	18,414	9,581	27,995	34%
4	20,386	7,608	27,994	27%
5	21,575	6,420	27,995	23%
6	22,729	5,266	27,995	19%
7	23,983	4,011	27,994	14%
8	25,047	2,948	27,995	11%
9	25,979	2,016	27,995	7%
10	26,839	1,154	27,993	4%
Total	213,019	66,928	279,947	24%

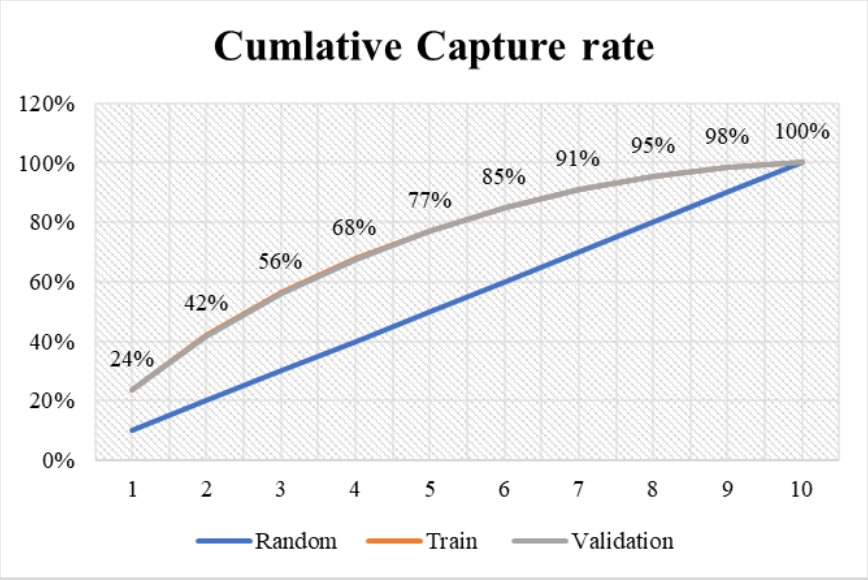

Fig 4. Cumulative Capture Rate: Train & Validation

Capture Rate Analysis:

- Top Decile: Captures 24% of responders, outperforming random selection.
- Top 3 Deciles: Cover 56% of total responders, balancing precision and coverage.
- Consistency: Similar capture rates for training and validation highlight robust model performance.

Key Takeaways:

- The model effectively targets high-propensity customers, optimizing marketing efforts.
- Stability across training and validation ensures scalability and reliability for future cohorts.
- Focused campaigns targeting the top deciles will yield maximum impact.

Back Test Results

Table 3 – Back Test Results

Back Test Results				
Decile	**Non-Response**	**Response**	**Total**	**Response Rate**
1	86,802	100,597	187,399	54%
2	116,275	71,125	187,400	38%
3	131,412	55,988	187,400	30%
4	139,502	47,898	187,400	26%
5	147,278	40,122	187,400	21%
6	155,984	31,416	187,400	17%
7	162,539	24,861	187,400	13%
8	169,395	18,005	187,400	10%
9	174,151	13,249	187,400	7%
10	179,355	8,044	187,399	4%
Total	1,462,693	411,305	1,873,998	22%

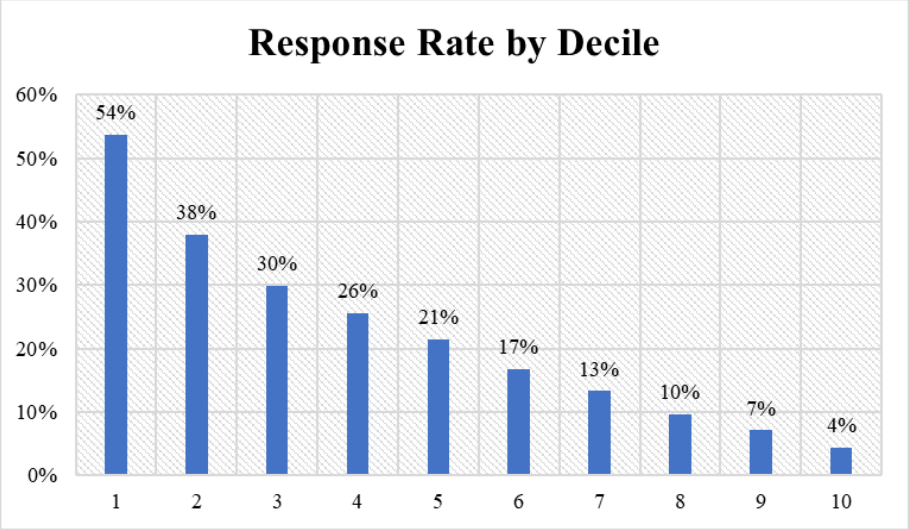

Fig 5. Response Rate by Decile (Back Test)

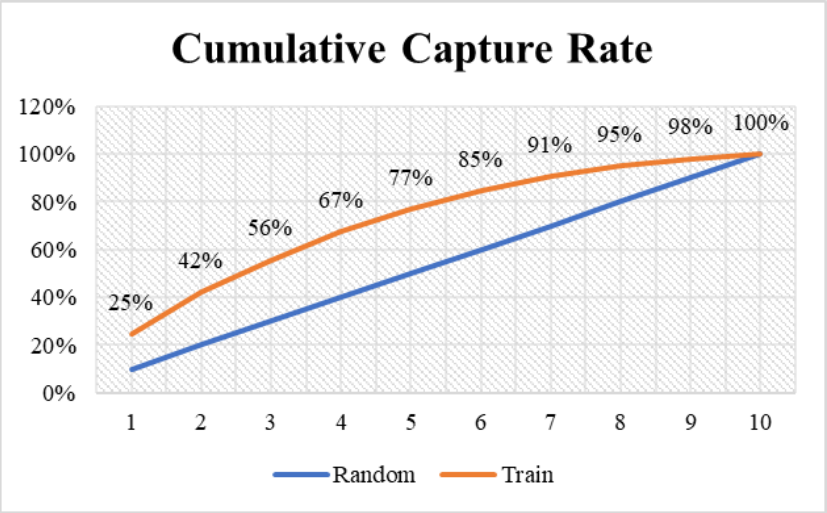

Fig 6. Cumulative Capture Rate (Back Test)

The model demonstrates strong performance on back-test data, maintaining trends consistent with training and validation datasets. The top 3 deciles successfully capture 56% of total responders, highlighting their effectiveness in prioritizing high-propensity customers**.**

4.6 Variables in the Model

Table 3 – Model Variables

Variable in the Model	Relative Importance
# Customer paid for service in the last 12 months	1.00
Average days between Repair orders	0.65
# Customer paid for service in all months	0.40
# Repair orders in the last 6 months	0.32
Months since the last repair order	0.31
Total customer payment in the last 12 months	0.31
%Engine related services performed	0.26
Flag indicating if the Customer has come for service in the last 12-24 months	0.26
Covid flag	0.02

The table above highlights the top predictors used in the model, with the "Sign" column indicating the direction of each variable's relationship with the propensity score. Notably, the highest-ranking predictors are primarily related to recency, frequency, and monetary (RFM) transactional behaviors, underscoring the importance of first-party service history in understanding customer engagement.

These RFM features were systematically engineered within the observation window defined for each cohort. Recency was measured as the number of days since the customer's last service event, frequency as the count of service interactions in the lookback period, and monetary as the total customer pay amount. By aligning these features to the cohort date, the model ensures that only historical, available data is used for prediction, enabling realistic, time-consistent scores that can be applied prospectively in campaign planning.

Beyond core transactional variables, the model incorporated customer identifiers (e.g., Household ID, VIN) to allow segmentation at granular levels and support personalized targeting. It also included categorical features such as service types and an economic context indicator, the Covid Flag. The Covid Flag was set to 1 during periods of significant market disruption and 0 otherwise, allowing the model to adjust

individual propensity scores to account for external conditions like the Covid-19 pandemic. This approach demonstrates how a cohort-based framework facilitates integrating both behavioral and macroeconomic signals, ensuring more accurate and context-aware predictions that align with business strategy.

5. FINDINGS & CONTRIBUTIONS

This study demonstrates the application of a cohort-based segmentation framework to predictive modeling in automotive aftersales marketing. The case study showed that structuring transactional data around cohort dates enabled the creation of meaningful temporal features (recency, frequency, monetary value) while avoiding data leakage.

Key empirical findings include:

- **Lift and Capture Performance**: The model achieved a lift of 2.7 in the top decile, capturing 24% of responders, and covered 56% of responders within the top three deciles, demonstrating strong prioritization of high-propensity customers.
- **Model Stability**: Consistent lift and capture rates across training, validation, and back-test cohorts confirm the framework's ability to generalize to unseen data.
- **Temporal Back-Testing**: Validation on a held-out cohort reinforced the importance of respecting temporal separation in predictive modeling to ensure realistic, actionable scoring.
- **Trend Analysis**: Response rates plotted across cohorts revealed clear patterns, including a dip during the COVID-19 lockdown period, highlighting how cohort-based setups can uncover external influences on customer behavior.

Contributions to Marketing Analytics Theory and Practice:

- The work extends cohort analysis theory beyond high-frequency contexts such as subscriptions or loyalty programs to low-frequency, high-value service transactions, addressing a noted gap in the literature.
- By enforcing observation and outcome windows, the approach operationalizes temporal consistency in feature engineering, aligning with best practices for causal inference and avoiding leakage in predictive modeling.
- The findings support the theoretical importance of RFM-based segmentation, demonstrating its value even in irregular purchase environments by systematically structuring historical service data.
- The study offers practical guidance for marketing analysts on integrating first-party transactional data with macroeconomic indicators (e.g., COVID-19 flag) to improve predictive accuracy.
- Finally, the framework provides a generalizable template for other industries with infrequent, discretionary service interactions (e.g., healthcare scheduling, equipment maintenance, B2B services).

6. LIMITATIONS

While cohort-based analysis provides valuable benefits for predictive modeling, it has limitations:

- **Noisy or Incomplete Data**: This approach relies heavily on historical records. Missing, inconsistent, or inaccurate data can bias engineered features, especially metrics like recency or frequency that depend on service or purchase history. For example, *if service visits are underreported in dealership systems, the model may mistakenly score engaged customers as inactive*.
- **Granularity vs. Complexity**: Choosing cohort granularity requires trade-offs. Very granular cohorts may lead to sparse data and unstable models, while overly broad cohorts risk masking meaningful differences in behavior. *For example, monthly cohorts might have too few events to train effectively, while yearly cohorts could blur seasonal buying patterns.*
- **Short Customer Lifecycles**: In industries with shorter customer lifecycles, compressed observation windows may limit the ability to derive meaningful patterns. *For example, in fast-moving retail or app-based services, customer behavior can change quickly, making long-term transaction history less predictive*.
- **Static Framework**: Fixed cohort timeframes may fail to capture evolving customer behaviors over time. *For instance, marketing strategies or external events like economic downturns can shift purchasing patterns in ways that require re-defining observation and performance windows or recalibrating the model*.

These challenges highlight the need for careful implementation and refinement to maximize the effectiveness of cohort-based analysis. Additionally, models built on cohort frameworks require ongoing monitoring and periodic recalibration to remain effective

as customer behaviors and market conditions evolve

7. CONCLUSION

Using cohort dates to define the eligible population base offers clear benefits for predictive modeling, especially in contexts with irregular, discretionary purchase patterns that have received limited attention in literature. By structuring data into observation and performance windows, cohort-based segmentation enables robust temporal feature engineering, trend analysis, and consistent data preparation even in low-frequency service settings. This approach improves model accuracy and generalizability while aligning predictive analytics with real-world marketing cycles. Cohort-based design also facilitates insights into event rates, seasonality, and evolving behavioral trends, making it applicable beyond automotive to sectors like retail, banking, healthcare, and subscription services. Future research could further extend this framework through real-time analytics, dynamic segmentation, personalized marketing integration, and incorporation of macroeconomic or external signals to enhance model relevance and precision.

References

1. B. Omidvar-Tehrani, S. Amer-Yahia and L. V. S. Lakshmanan, "Cohort Representation and Exploration," 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), Turin, Italy, 2018, pp. 169-178, doi: 10.1109/DSAA.2018.00027

2. Lemmens, Aurélie & Croux, Christophe. (2006). Bagging and Boosting Classification Trees to Predict Churn. Journal of Marketing Research. 43. 10.1509/jmkr.43.2.276.

3. Blattberg, Robert & C., Robert & Byung-Do, & Kim, & Neslin, Scott & A., Neslin. (2008). Database Marketing: Analyzing and Managing Customers.

4. Gupta, Sunil & Zeithaml, Valarie. (2006). Customer Metrics and Their Impact on Financial Performance. Marketing Science. 25. 718-739. 10.1287/mksc.1060.0221.

5. Neslin, Scott & Grewal, Dhruv & Leghorn, Robert & Shankar, Venkatesh & Teerling, Marije & Thomas, Jacquelyn & Verhoef, Peter. (2006). Challenges and Opportunities in Multichannel Customer Management. Journal of Service Research - J SERV RES. 9. 95-112. 10.1177/1094670506293559.

6. Chen, J., Chen, M., Liao, W. and Chen, T. (2009), "Influence of capital structure and operational risk on profitability of life insurance industry in Taiwan", Journal of Modelling in Management, Vol. 4 No. 1, pp. 7-18. https://doi.org/10.1108/17465660910943720

7. C. X. Ling and C. Li, “Data Mining for Direct Marketing: Problems and Solutions,” Proceedings of International Conference on Knowledge Discovery from Data (KDD 98), New York City, 27-31 August 1998, pp. 73-79.

8. Larsen, Nicholas & Stallrich, Jonathan & Sengupta, Srijan & Deng, Alex & Kohavi, Ron & Stevens, Nathaniel. (2023). Statistical Challenges in Online Controlled Experiments: A Review of A/B Testing Methodology. The American Statistician. 78. 1-32. 10.1080/00031305.2023.2257237.

9. Neslin, S.A., Gupta, S., Kamakura, W., Lu, J.X. and Mason, C.H. (2006) Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models. Journal of Marketing Research, 43, 204-211.

10. Brownlee, J. (2020). Probability for Machine Learning: Discover How To Harness Uncertainty With Python. San Francisco: Machine Learning Mastery.

11. Tummalapalli, Vaibhav. (2025). Stratified sampling in Cohort-based data for Machine learning Model development. International Scientific Journal of Engineering and Management. 04. 1-8. 10.55041/ISJEM03377

12. V. Tummalapalli and K. Konakalla, "Statistical Techniques for Feature Selection in Machine Learning Models," International Journal for Innovative Research in Multidisciplinary Pursuit and Studies (IJIRMPS), vol. 13, no. 3, pp. 1-8, 2025, doi: 10.37082/IJIRMPS.v13.i3.232566

13. Chen, Yen-Liang & Kuo, Mi-Hao & Wu, Shin-yi & Tang, Kwei. (2009). Discovering recency, frequency, and monetary (RFM) sequential patterns from customers’ purchasing data. Electronic Commerce Research and Applications. 8. 241-251. 10.1016/j.elerap.2009.03.002.

14. Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert Systems with Applications, 36(2), 2592–2602. https://doi.org/10.1016/j.eswa.2008.02.021

15. Hughes, A. M. (1996). The complete database marketer: second-generation strategies and techniques for database marketing. McGraw-Hill

16. V. Tummalapalli, “Feature Engineering for Building Machine Learning Models in Automotive Industry,” International Scientific Journal of Engineering and Management, vol. 4, no. 8, pp. 1–9, 2025. doi: 10.55041/ISJEM04985.

17. V. Tummalapalli, “Comprehensive study of data imputation techniques for machine learning models,” International Journal of Innovative Research in Engineering & Multidisciplinary Physical Sciences, vol. 13, no. 4, 2025, doi: 10.37082/IJIRMPS.v13.i4.232674.

18. V. Tummalapalli, “Machine learning pipeline for automotive propensity models,” International Journal of Core Engineering & Management, vol. 8, no. 3, 2025, ISSN 2348-9510

19. V. Tummalapalli, “Understanding distance metrics in KNN imputation: Theoretical insights and applications,” Journal of Mathematical & Computer Applications, vol. 4, no. 4, pp. 1–4, 2025. doi: 10.47363/JMCA/2025(4)208.

20. Vaibhav Tummalapalli. (2025). Outlier Detection & Treatment for Machine Learning Models. International Journal of Innovative Research and Creative Technology, 11(3), 1–8. https://doi.org/10.5281/zenodo.16500050