ISSN: 1813-4890

Theoretical Research on the Full-Chain Data Attribution Model in the Automotive Industry

Jianshu Li, Lin Zhang, Rui Shi*
CATARC Brand Technology (Tianjin) Co., Ltd., China
Corresponding author: shirui2022@catarc.ac.cn

Abstract

The domestic automotive market has transitioned from an expansion phase to a stage of stock competition, with increasingly complex marketing scenarios and consumer purchase decisions becoming more online-oriented and fragmented. Traditional lead-driven marketing models are constrained by data silos, making it difficult to accurately quantify the true value of each marketing touchpoint, resulting in inefficiencies in budget allocation and low conversion returns—key industry challenges. Drawing on information diffusion theory, consumer decision journey models, and causal inference methods, this study constructs a full-funnel data attribution model for the automotive industry， covering content, leads, and conversions. By integrating multi-source data from social media engagement, private domain operations, and transaction conversions, and leveraging tools such as Markov chains and survival analysis, the model enables comprehensive tracking of user behavior across the entire lifecycle and quantifies channel contributions. Using the marketing practices of joint-venture automakers as a case study, we validate the model's effectiveness and explore technical pathways for enabling attribution analysis through digital intelligence platforms. This research provides theoretical and methodological support for precise marketing decisions in the automotive sector and offers a novel analytical framework for quantitative studies in mass consumer goods marketing.

Keywords

Full-funnel attribution; automotive marketing; data silos; Markov chain; user lifecycle.

1. Introduction

China's automotive industry is currently undergoing profound transformation, and consumers' car-purchasing decision-making processes have fundamentally changed. Buyers no longer passively receive one-way marketing messages from automakers but instead actively compare information, conduct word-of-mouth research, and gradually build trust through online channels such as short videos, live streaming, social media platforms, and vertical automotive websites. Against this backdrop, effective brand-user engagement has become increasingly scarce, and attributing marketing effectiveness has grown significantly more complex. However, the current evaluation system for automotive marketing lags behind evolving consumer behaviors. For years, automakers and dealers have relied on a linear funnel model—"lead generation → phone invitation → in-store conversion"—measuring marketing success solely by lead volume and call response rates. This coarse approach has clear limitations: consumers typically experience dozens of marketing touchpoints over an extended period before making a purchase; there are nonlinear, cross-channel synergies between online exposure, private domain operations, and offline store visits, making traditional last-click attribution methods highly likely to undervalue the contribution of earlier marketing interactions.

Attribution analysis models can address the aforementioned issues, and attribution analysis has already been applied across various domains. For instance, reference [1] proposes a method for predicting injury risk and identifying causative factors in dynamic driving accidents of autonomous vehicles (AVs), based on CTGAN and XGBoost, to tackle the problem of high proportions of stationary samples in existing AV accident datasets. Based on 991 accident reports from NHTSA, a high-dimensional feature set incorporating vehicle characteristics, collision objects, and road environments is constructed. A CDQS quality evaluation system is used to select optimal CTGAN configurations for augmenting minority class samples, thereby alleviating data imbalance. An XGBoost prediction model is built, with results cross-validated using SHAP values and text mining. Results show that the CTGAN-XGBoost model achieves the best generalization performance; speed, accident location, and pre-collision status are identified as core risk factors, with injury risk sharply increasing when speeds exceed 20 km/h, and higher risks observed at intersections and during turning maneuvers. Accident risk exhibits significant heterogeneity—daytime conditions, collisions involving vulnerable road users, aggressive ODD strategies in urban areas, and high-speed traffic flow under clear weather all elevate injury risk. Reference [2] addresses the challenges of deteriorating mobile network quality, declining user experience, and rising operational costs by integrating terminal, base station, and meteorological data. Using four months of real-world measurements from a city-wide network, a heterogeneous dataset is constructed, and a random forest-based multi-decision tree voting approach is employed for 5G network quality degradation attribution analysis. The study uses RSRP and SINR as quality assessment criteria, achieving an accuracy of 74% and an AUC of 0.75 in urban tests. The findings confirm that spectrum allocation, user load, and weather conditions are key influencing factors. The model enables visual identification of root causes behind poor-performing cells, providing quantitative support for network optimization and operations.

Automotive marketing attribution is constrained by long decision cycles, fragmented data, and disconnected online-offline customer journeys, creating significant theoretical and practical barriers in model development. To address these challenges, this paper proposes a comprehensive end-to-end data attribution model tailored for the automotive industry. Theoretically, this study expands the applicability of attribution theory to high-value, offline-intensive, and delayed-consumption scenarios. Unlike fast-moving consumer goods, which follow a straightforward and rapid conversion path, automotive purchase decisions involve numerous offline touchpoints and cross-cycle delays, demanding more sophisticated time-series modeling and causal inference. The proposed "content–lead–conversion" framework offers a new paradigm for attributing complex automotive consumer decisions. Practically, the model aligns with the industry's digital transformation needs, enabling automakers to clarify the causal relationship between marketing investments and business outcomes, optimize cross-channel budget allocation, and enhance customer acquisition efficiency. A case study at Changan Automobile demonstrates that digital end-to-end marketing operations can effectively improve lead conversion rates and drive new sales growth, proving highly valuable in real-world implementation.

2. Automotive Marketing and Attribution Theory Foundations

2.1. The Dilemma of Automotive Marketing

The uniqueness of the automotive industry's attribution model stems from the deep structure of its marketing system. The complexity in marketing arises not only from vehicles' high purchase value but also from a distinctive channel architecture and data ecosystem, which manifests primarily in three forms of fragmentation. First, organizational fragmentation: functions such as brand communication, digital advertising, lead management, and retail sales

ISSN: 1813-4890

operate independently, each with its own KPIs and data systems. While they perform their respective roles, there is a lack of unified measurement standards and collaborative mechanisms, resulting in disconnection at the strategic marketing level. Second, channel fragmentation: automakers handle comprehensive brand exposure, while dealerships manage offline conversions. Data between these two parties cannot be integrated, leaving user touchpoints across multiple channels isolated and making it difficult to reconstruct the complete car-buying journey. Third, online-offline fragmentation: the core transaction process occurs offline, disconnected from online marketing data. Traditional last-click attribution significantly undervalues the impact of earlier online interactions. These overlapping fragmentations have led the automotive marketing sector into a dilemma of high investment and low visibility, making it impossible to clearly trace the full customer journey from awareness to conversion. A full-funnel attribution model can bridge these broken marketing value chains by integrating data, effectively addressing the industry's attribution challenges.

2.2. Attribution Analysis Theory

Marketing attribution is fundamentally a causal inference problem in multi-touchpoint scenarios, aiming to scientifically trace conversion outcomes back to their preceding marketing touchpoints. Both academia and industry have developed various mature attribution methods. Single-touchpoint attribution includes basic models such as last-click, first-click, and linear attribution, which are simple to calculate and easy to implement but suffer from significant drawbacks—fragmenting the user's complete decision-making journey and often underestimating the marketing value of early brand awareness touchpoints. Multi-touchpoint attribution models overcome these limitations by fairly allocating conversion credit across multiple channels. These models primarily fall into three categories: regression-based, probabilistic, and time-weighted approaches. They quantify channel contributions by modeling conversion probabilities, calculating conditional probabilities, or incorporating time-decay weighting, respectively. Empirical studies in the auto insurance industry have validated the effectiveness of these methods. Markov chain attribution has emerged as a mainstream approach[3-4] in recent years, modeling the user's consumption journey as a state transition process. It quantifies marginal contribution through touchpoint removal effects, effectively capturing temporal dependencies among touchpoints and cross-channel interactions, making it well-suited for complex conversion scenarios.

2.3. Special Challenges in Attribution Research for the Automotive Industry

Translating the general attribution theory framework into automotive marketing scenarios requires addressing three unique challenges. First is the issue of temporal decay: car purchase decisions span weeks to months, meaning early marketing touchpoints exhibit long-term lag effects, necessitating precise alignment of time windows and decay functions in the model. Second is data sparsity and the curse of dimensionality: while car sales data is limited, the dimensions—such as marketing channels, content combinations, and competitor variables—are complex, making high-dimensional modeling prone to overfitting and estimation bias. Third is the blind spot in offline behavior observation: critical offline decision-making actions such as in-store test drives and word-of-mouth recommendations are difficult to fully capture, leading to potential misestimation of online channel contributions. Therefore, building an effective attribution model for the automotive industry requires a balanced approach that integrates scenario-adapted algorithmic frameworks, robust data governance mechanisms, and strong platform support capabilities.

3. Theoretical Framework of the Full-Funnel Attribution Model

3.1. Attribution Logic

This paper proposes a three-stage attribution framework—Content (C), Lead (L), and Deal (D)—as the core analytical architecture for end-to-end automotive marketing attribution. The logic behind this framework stems from a breakdown of the consumer decision-making process in the automotive industry. The standard consumer decision journey can be divided into four stages: awareness, consideration, intent, and purchase. Mapping this theoretical framework onto observable data dimensions in automotive marketing, as shown in Table 1, enables practical application and analysis.

Table 1 Observable Data Dimensions

Serial Number	Data dimensions	Data Subject	Note
1	Content	Impressions, engagement, content completion rate, and follower growth	Consumers discover brands through social media content, official brand information, KOL recommendations, and other channels, forming initial awareness and interest.
2	Lead	Lead volume, lead source channels, lead quality score, response timeliness	Consumers can engage through test drive forms, calling the 400-number hotline, adding the company's WeChat account, or participating in online activities.
3	Deal	Visit rate, test drive rate, conversion rate, transaction amount	Consumers complete conversion actions such as visiting the store, test driving, signing contracts, and vehicle delivery.

3.2. Variable System and Data Fusion Framework

This paper designs a three-tier variable system based on the "Content (C)-Lead (L)-Deal (D)" framework, comprising content-level variables, lead-level variables, and conversion-level variables. The content-level variables include dissemination data such as publishing channels, content formats, marketing themes, influencer persona, posting frequency, exposure, and engagement rates. The lead-level variables encompass lead source, acquisition time, user profile information, response timeliness, WeCom communication, test drive invitations, and other nurturing behavior data. The conversion-level variables serve as core outcome indicators, covering key conversion metrics such as whether a sale was completed, the model sold, price, conversion timing, and financing options for vehicle purchase, thereby providing comprehensive variable support for end-to-end attribution modeling.

The core of multi-source data integration lies in establishing a unified user ID system to connect user behavior data from fragmented systems such as social media, WeCom (Enterprise WeChat), CRM, and DMS, thereby forming a complete user access journey. This process leverages the capabilities of a CDP data platform to standardize, clean, map, and merge data from multiple sources, achieving unified data integration.

4. Model Development

4.1. Markov-based Attribution Model

The Markov chain model treats the consumer journey as a stochastic process in a finite state space. $S=\{s_1,s_1,\cdots,s_n,t\}$ Let the state space be S, and the consumer's path consists of a sequence of state transitions, each satisfying the first-order Markov property[3-4], which is mathematically expressed by Equation (1).

ISSN: 1813-4890

$$P(s_{k+1}=j \mid s_1, s_2, \cdots, s_k) = P(s_{k+1}=j \mid s_k) \tag{1}$$

Among them, M is the transition probability matrix, and the mathematical expression of M is $M_i^j = P(s_{k+1}=j \mid s_k = i)$. The estimation of the conversion probability is obtained by calculating the absorption probabilities of absorbing states. The core operation of attribution is to compute the removal effect for each channel state s_i, and the expression that defines the decrease of p_t in the overall conversion probability after removing state s_i from the state space is given by Equation (2).

$$RE_i = P_t(S) - P_t(S \mid \{s_i\}) \tag{2}$$

Here, $P_t(S)$ represents the conversion probability under the complete state space, and $P_t(S \mid \{s_i\})$ represents the conversion probability after removing channel i; a larger removal effect indicates that the channel plays a more critical role in final conversion. By normalizing the removal effects, we obtain the attribution weights for each channel, as expressed mathematically in Equation (3).

$$w_i = \frac{RE_i}{\sum_{i=1}^{m} RE_i} \tag{3}$$

Markov chain models effectively capture the temporal dependencies and dynamic relationships among multiple marketing touchpoints, precisely aligning with automotive users' long-term, sequential purchase decision processes. They accurately depict consumers' cross-channel, stage-wise cognitive and conversion evolution, making them better suited to the complex attribution scenarios in automotive marketing.

4.2. Time Decay Function and Survival Analysis Model

The decision-making cycle for car purchases is long, making it crucial to properly model temporal factors. This paper employs two approaches—time-decay weighting and survival analysis models—to handle the time dimension.

4.2.1. Time Decay Function

The time decay function in this paper adopts an exponential decay form, whose mathematical expression is given by Equation [5-6](4).

$$f(\Delta t_i) = e^{(-\alpha \cdot \Delta t_i)} \tag{4}$$

From Equation (4), $f(\Delta t_i)$ is the contribution weight adjusted through the decay function, Δt_i is the time difference between contact point i and the conversion occurrence; α is the decay rate parameter, whose value range is $\alpha > 0$, which needs to be calibrated according to specific product category characteristics. Generally, the value of α for the automotive industry should be lower than that for fast-moving consumer goods, reflecting its longer decision-making cycle. Therefore, the revised mathematical expression for channel contribution is given by Equation (5).

$$f(w_i) = \frac{w_i \cdot f(\Delta t_i)}{\sum_{i=1}^{m} w_i \cdot f(\Delta t_i)} \tag{5}$$

4.2.2. Survival Analysis Model

Survival analysis complements traditional attribution research by introducing a temporal dimension, focusing on the time distribution patterns from users' first marketing touchpoint to final purchase conversion. It quantifies the impact of various channels and marketing characteristics on the duration of consumers' car-buying decision process, offering a novel time-series analytical approach to understanding long-cycle automotive consumption decisions. The hazard function for consumers from initial contact to completed purchase is given by Equation (6).

$$h(t)=\lim_{\Delta t\to 0}\left(\frac{P(t\le T+\Delta t\mid T\ge t)}{\Delta t}\right) \tag{6}$$

From Equation (6), T is the duration from a consumer's first contact to completion of purchase, with other parameters as described above. Additionally, the Cox proportional hazards model is employed to estimate the impact of channel characteristics on conversion rates, with the mathematical expression of the Cox proportional hazards model given by Equation (7).

$$h(t)=h_0(t)\cdot e^{\beta_1 x_1+\beta_2 x_2+\cdots+\beta_t x_t} \tag{7}$$

From Equation (7), x_t denotes the channel characteristic variables such as contact with a specific channel and contact frequency, and β_t is the corresponding risk coefficient. If $\beta_t>0$, it indicates that this channel characteristic can accelerate the conversion decision, and this information provides supplementary value for understanding the channel's contribution to consumer decision-making efficiency.

5. Research Findings and Future Perspectives

5.1. Research Findings

This study addresses the challenges in automotive marketing attribution by constructing a full-funnel attribution model based on "content--**lead**--**deal**" , integrating Markov chains and survival analysis to quantify user lifecycle behaviors and fairly allocate channel contributions. The research reveals that consumer decision-making follows a three-stage pattern, with significantly different dominant channels across stages. Last-click attribution severely underestimates the value of early awareness channels, leading to misallocation of marketing budgets. Moreover, successful implementation of the model depends not only on advanced algorithms but also on systematic support including data infrastructure, engineering tools, and organizational safeguards. Breaking down data silos, enabling business-oriented interpretation of attribution results, and establishing collaborative mechanisms between manufacturers and dealers are critical for translating attribution from theory into practice.

5.2. Future Research Directions

This study still has certain limitations: the empirical analysis focuses solely on online marketing channels, excluding offline touchpoints such as in-store experiences and word-of-mouth referrals, which may overestimate the contribution of online marketing; the first-order state assumption in the Markov chain model is relatively simplified and fails to fully account for the long-term effects of multiple touchpoints accumulated earlier; and the sample is limited to a single regional market, so the model's external generalizability across regions and brands remains to be validated.

Future research can be deepened in three aspects: integrating offline data such as store foot traffic and test drive records to build an integrated online-offline, full-scenario attribution model; adopting deep reinforcement learning methods to advance the model from post-hoc attribution evaluation to real-time marketing budget optimization decisions; and leveraging compliant technologies like multi-party secure computing to enable secure cross-organizational data fusion, thereby expanding the model's data dimensions. The automotive industry has now entered a phase of stock competition, where refined marketing has become a core competitive advantage. A full-funnel attribution model provides a scientific framework for evaluating marketing effectiveness, helping automotive marketing evolve toward greater precision and intelligence.

ISSN: 1813-4890

References

[1] Wang, X. F., Xu, J., Zheng, P., et al. Identification of key causes of injuries in dynamic driving accidents for autonomous vehicles based on CTGAN-XGBoost and cross-modal attribution. China Journal of Highway and Transport. https://link.cnki.net/urlid/61.1313.U.20260526.1004.004

[2] Wen, C. J., Ye, M. H., & Zhao, Y. Network quality degradation attribution analysis based on multi-source data fusion and explainable random forest. Telecommunications Engineering Technology and Standardization, 2026, 39(4), 1-9. https://doi.org/10.13992/j.cnki.tetas.2026.04.014

[3] Cao, Z. J., Chen, X. Z., Liu, T., et al. Intelligent learning and efficient zoning simulation method for Markov random field stratigraphic models based on position information matrix. China Journal of Highway and Transport. https://link.cnki.net/urlid/61.1313.u.20260709.0941.002

[4] Xue, L. L., Wang, Y. Z., Bo, T., et al. Research on behavior decision method for tractor-trailers based on Markov decision process. Automotive Practical Technology, 2026, 51(13), 26-32. https:// doi.org/ 10.16638/j.cnki.1671-7988.2026.013.005

[5] Wang, X. F. A study on user interest changes in academic virtual communities based on time decay functions. Information Exploration, 2022, (9), 111-116.

[6] Liu, S. N. Research and application of collaborative filtering algorithm based on time decay and dimensionality reduction optimization [Doctoral dissertation]. Nanchang University. https:// doi.org/ 10.27232 /d.cnki. gnchu. 2022.003638