INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 21, Issue: 5, Month: October, Year: 2026
Article Number: 7431, https://doi.org/10.15837/ijccc.2026.5.7431

# Multi-Source Heterogeneous Data Fusion for Predicting Customer Lifetime Value in the New Energy Vehicle Market

L. Mei, Y. Shang, Y. Deng, L. Fan, Z. Zhang, Y. Yang

**Lei Mei**
Business School of Geely University
Eastern New District, Chengdu, Sichuan Province, 641423, China
Farabi international Business School, Al-Farabi Kazakh National University, Kazakhstan
meilei@guc.edu.cn

**Yangyang Shang**
Research Institute of Innovation Economy named after Sh. Musakozhoev
Sh. Musakozhoev, Kyrgyz Republic, 720033, Bishkek, Kyrgyz Republic
jscjcxcyxy@163.com

**Yuqin Deng***
Business School of Geely University
Eastern New District, Chengdu, Sichuan Province, 641423, China
*Corresponding author: dengyuqin@guc.edu.cn

**Lifei Fan**
Geely Automobile Group Co., Ltd.
No. 1760 Jiangling Road, Binjiang District, Hangzhou, Zhejiang Province, 310051, China

**Zeming Zhang**
Business School of Geely University Eastern New District, Chengdu, Sichuan Province, 641423, China
zemingzhang@guc.edu.cn

**Yang Yang**
Business School of Geely University
Eastern New District, Chengdu, Sichuan Province, 641423, China
yangyang@guc.edu.cn

**Abstract**

Accurate prediction of customer lifetime value (CLV) is essential in the new energy vehicle (NEV) market, where long-term profitability increasingly depends on charging-related services and infrastructure interaction rather than one-time vehicle sales. However, NEV customer behavior is influenced by heterogeneous and dynamically evolving factors, including charging usage patterns, spatial infrastructure availability, and regional market conditions, which challenge conventional transaction-based CLV models.This study proposes a multi-source heterogeneous data fusion framework for NEV CLV prediction using exclusively publicly available and verifiable data. The framework integrates user-level charging transaction records with infrastructure interaction features through a modality-aware fusion architecture, in which behavioral and contextual information is encoded and aligned along temporal and spatial dimensions. CLV is estimated under a discounted usage-based revenue formulation, and a fully reproducible experimental protocol is established for evaluation.The results demonstrate that heterogeneous data fusion improves both predictive accuracy and value-based customer ranking, offering practical decision support for NEV charging service providers and ecosystem stakeholders.

**Keywords:** Customer Lifetime Value, New Energy Vehicles, Heterogeneous Data Fusion, Electric Vehicle Charging Behavior, Machine Learning.

# 1 Introduction

Customer lifetime value (CLV) has long been recognized as a central construct for linking individual customer behavior with long-term firm performance. In many application domains, however, customer relationships are non-contractual, and future value cannot be inferred from explicit churn events or renewal decisions. Instead, CLV must be estimated from observed behavioral traces, such as usage frequency, consumption intensity, and temporal regularity, which only indirectly reflect customers' latent continuation intentions [1, 2]. This characteristic makes CLV prediction inherently uncertain, particularly when the forecasting horizon extends beyond the immediate future. As a result, CLV has been widely adopted in marketing, operations, and decision support research as a tool for guiding resource allocation, customer prioritization, and service differentiation under uncertainty [23]. Prior studies have shown that even coarse estimates of long-term customer value can substantially improve strategic decision making compared with short-term or transaction-level metrics. However, the majority of existing CLV models focus on relatively homogeneous data representations, often relying on aggregated transaction histories or simplified behavioral summaries, despite empirical evidence that incorporating behavioral regularity and usage dynamics can substantially improve customer value estimation [24].

These limitations are especially pronounced in usage-based service systems such as electric vehicle (EV) charging networks. In this context, charging demand is not determined solely by individual preferences or historical consumption patterns. Instead, user behavior is strongly conditioned by the accessibility, spatial distribution, and operational characteristics of charging infrastructure, which impose external constraints on when and where charging sessions can occur [25, 26]. Consequently, two users with similar historical charging volumes may exhibit markedly different future value trajectories depending on their interaction with the underlying infrastructure. Ignoring such contextual factors can therefore lead to biased or unstable CLV estimates in energy service ecosystems.

Another challenge in usage-based CLV prediction concerns robustness across forecasting horizons. Forecasting accuracy is known to deteriorate as the prediction horizon increases due to the accumulation of uncertainty and the growing influence of unobserved factors [27]. While short-term forecasts may benefit primarily from recent behavioral signals, medium- and long-term predictions require models to remain informative under increasingly sparse and volatile future observations. Despite its practical importance, systematic evaluation of CLV robustness across multiple forecasting horizons remains relatively underexplored, particularly in infrastructure-dependent service environments.

From a broader methodological perspective, recent advances in data-driven modeling emphasize the value of integrating heterogeneous information sources to capture complementary dimensions of economic behavior [28]. Rather than treating behavioral, contextual, and structural descriptors as interchangeable inputs, heterogeneous data fusion frameworks aim to exploit their distinct informational roles. Motivated by these considerations, this study proposes a multi-source feature fusion approach for CLV prediction in the EV charging context and evaluates its performance across multiple forecasting horizons. By explicitly examining robustness and feature contribution patterns, the study seeks to clarify not only whether heterogeneous data improve prediction accuracy, but also how and under what conditions such improvements arise.

# 2 Literature Review

## 2.1 Customer Lifetime Value Modeling

Customer lifetime value (CLV) has long been recognized as a fundamental construct for assessing long-term customer profitability and guiding strategic resource allocation. Early CLV research was primarily grounded in transaction-based statistical modeling, where future value was inferred from historical purchase behavior using indicators such as recency, frequency, and monetary value [1]. Probabilistic approaches further formalized customer behavior under non-contractual settings by explicitly modeling repeat purchasing and attrition processes, with the buy-till-you-die family of models representing a seminal contribution in this stream [2, 3]. With the proliferation of digital transaction records and behavioral data, machine learning techniques have been increasingly adopted for CLV prediction. Tree-based models, ensemble learning, and neural networks have demonstrated improved predictive accuracy compared with traditional parametric approaches, particularly in environments characterized by nonlinear consumption patterns and heterogeneous customer segments [4]-[7]. These methods relax restrictive distributional assumptions and allow complex interactions among behavioral variables to be captured more effectively. Despite these advances, most CLV studies remain focused on single-source transactional data and implicitly assume relatively stable and homogeneous consumption environments. Such assumptions are increasingly challenged in emerging service ecosystems, where customer value is shaped by irregular usage patterns and external constraints. Recent research has therefore begun to emphasize that CLV formation depends not only on individual purchasing behavior but also on contextual and environmental factors, such as service accessibility, technological infrastructure, and market conditions [8, 9]. This shift has motivated the development of multi-dimensional CLV modeling frameworks. However, empirical applications remain limited, particularly in technology-driven and infrastructure-dependent markets, where usage-based value realization dominates over traditional purchase cycles.

### 2.2 Heterogeneous Data Fusion and Predictive Analytics

Heterogeneous data fusion refers to the systematic integration of information from multiple sources that differ in structure, scale, and semantic meaning, with the objective of improving inference quality and decision support. In computer science and engineering, data fusion techniques have been extensively studied in sensor networks and distributed systems, where structured association, uncertainty management, and stability are critical concerns [10]. A number of studies have contributed important methodological insights to this stream. Chen and Chung [11] proposed a systematic data fusion methodology emphasizing stability and convergence in multi-source environments. Subsequent research demonstrated that feature engineering combined with ensemble-based fusion strategies can significantly enhance predictive performance in decision-support and consumer analytics applications [12]. More recent work has extended fusion techniques to energy- and mobility-related domains, showing their effectiveness in extracting structured knowledge from heterogeneous behavioral and contextual data [13]. Heterogeneous data fusion has also attracted increasing attention in marketing analytics, recommender systems, and risk modeling. Empirical evidence consistently shows that integrating behavioral signals with contextual information leads to more robust and interpretable predictions than relying on a single data modality [14]-[16]. These findings suggest that data fusion is particularly well suited for socio-technical systems, where user behavior is jointly shaped by individual preferences and external constraints. Nevertheless, in many applications, fusion is still operationalized as simple feature aggregation, without explicit consideration of modality heterogeneity or alignment, limiting its explanatory and predictive potential.

### 2.3 Data-Driven Research in the New Energy Vehicle Domain

The rapid expansion of the new energy vehicle (NEV) market has stimulated extensive data-driven research. A substantial body of literature focuses on charging demand forecasting, load management, and infrastructure planning, leveraging high-frequency charging session data to optimize system-level performance and grid stability [17]-[19]. These studies provide valuable operational insights but typically abstract away from individual-level value assessment and long-term customer profitability. Parallel research has examined user perceptions and experiences in the NEV ecosystem using text mining and sentiment analysis. Studies based on online reviews and social media data show that concerns related to charging convenience, cost, and service reliability play an important role in shaping user behavior [13, 20]. While these analyses offer complementary behavioral perspectives, they are rarely linked to monetized value metrics or integrated into CLV prediction frameworks. An important limitation of existing NEV-related empirical studies lies in data accessibility. Many analyses rely on proprietary datasets obtained from manufacturers or charging service operators, which restrict reproducibility and cross-study comparability. In contrast, publicly available datasets—such as open charging session records, charging station registries, and regional vehicle adoption statistics—provide new opportunities for transparent and replicable research [21, 22]. However, existing work seldom exploits these open data sources in an integrated manner to model long-term customer value, leaving a significant methodological gap.

### 2.4 Research Gap

The above review highlights three interrelated gaps in the literature. First, CLV modeling in the NEV context remains underdeveloped and insufficiently adapted to usage-based value generation mechanisms driven by charging behavior. Second, although heterogeneous data fusion has demonstrated strong potential in predictive analytics, its application to CLV modeling using public NEV data is still limited, and modality heterogeneity is often inadequately addressed. Third, there is a lack of reproducible empirical frameworks that jointly consider behavioral, infrastructural, and contextual factors in NEV customer value assessment.This study addresses these gaps by proposing a multi-source heterogeneous data fusion framework for NEV CLV prediction based exclusively on publicly available datasets. By integrating charging behavior with infrastructure interaction features and market context in a transparent and reproducible manner, the study contributes to both the CLV literature and the broader research on data fusion and decision support systems.

## 3 Data Sources and Preprocessing

### 3.1 Data Selection Principles

To ensure transparency, reproducibility, and verifiability, this study strictly adheres to three data selection principles.First, all datasets employed in the analysis are publicly accessible through official repositories, peer-reviewed data journals, or government-supported open platforms. Second, the data sources provide sufficient documentation regarding data collection procedures, variable definitions, and update frequency, enabling independent replication. Third, the selected datasets collectively cover user-level behavioral signals, infrastructure-level constraints, and market-level contextual factors, which are essential for modeling customer lifetime value (CLV) in the new energy vehicle (NEV) ecosystem.Based on these principles, the empirical analysis integrates multiple open datasets spanning charging transactions, charging infrastructure characteristics, and regional adoption indicators. No proprietary, simulated, or manually fabricated data are used at any stage of the study.

### 3.2 Charging Transaction Data

User-level charging behavior constitutes the core behavioral signal for CLV estimation in the NEV market. This study relies on publicly released electric vehicle charging transaction datasets reported in peer-reviewed data journals and open research portals. These datasets typically record individual charging sessions with timestamps, energy consumption, charging duration, station identifiers, and, where available, transaction price information [4]. Charging transaction data provide a direct and monetizable proxy for customer engagement in the NEV ecosystem. Compared with traditional purchase data in retail settings, charging sessions exhibit irregular temporal patterns and substantial heterogeneity across users. Such characteristics make them particularly suitable for evaluating long-term value accumulation but also require careful preprocessing to address sparsity and temporal misalignment. In this study, raw charging records are first filtered to remove incomplete or invalid sessions, such as entries with missing timestamps or non-positive energy values. All timestamps are standardized to a unified format and aggregated to consistent temporal units (e.g., monthly intervals) to facilitate downstream CLV computation and model training. User identifiers are anonymized in the original datasets and are used solely for analytical grouping purposes.

### 3.3 Charging Infrastructure Data

Charging infrastructure availability plays a critical role in shaping charging behavior and long-term customer value. To capture infrastructure-related constraints and opportunities, this study incorporates publicly available charging station location and attribute data from national open data platforms [5].Infrastructure datasets typically include station geographic coordinates, charger types (e.g., Level 2, DC fast charging), operational status, and commissioning dates. These attributes enable the construction of spatial accessibility and capacity indicators that reflect the charging environment faced by individual users. Compared with transaction data, infrastructure data evolve more slowly over time but exert persistent effects on user behavior.For preprocessing, station-level data are cleaned to retain operational stations and to standardize geographic information. Spatial joins are then performed to associate charging sessions with nearby infrastructure characteristics. Infrastructure features are aggregated at the regional and temporal levels consistent with the charging transaction data, ensuring compatibility across data sources.

### 3.4 Market Adoption and Contextual Data

Beyond individual behavior and infrastructure conditions, regional market context influences long-term NEV usage intensity and customer retention. To capture these effects, the analysis incorporates publicly available electric vehicle registration and adoption statistics released by open data initiatives and governmental agencies [6].Market-level indicators include the number of registered electric vehicles, growth rates of NEV adoption, and regional penetration measures. These variables serve as contextual covariates reflecting policy support, market maturity, and network effects within a given region. Although such indicators do not directly represent individual behavior, they provide essential background information that conditions charging demand and long-term engagement.Market context data are aligned temporally with charging transaction records through interpolation or aggregation when reporting frequencies differ. All transformations are explicitly documented to avoid introducing artificial trends or biases.

### 3.5 Data Alignment and Feature Integration

A major challenge in multi-source heterogeneous data analysis lies in aligning datasets with different temporal resolutions, spatial references, and semantic meanings. In this study, data alignment is conducted along two primary dimensions: time and space.Temporal alignment is achieved by aggregating all data sources to a common time unit, enabling consistent longitudinal analysis of customer behavior and contextual conditions. Spatial alignment is performed by mapping charging stations and transactions to standardized regional units, such as cities or states, which also correspond to the reporting units of market adoption data.After alignment, features from different sources are integrated into a unified analytical dataset. Charging transaction features capture user-level behavioral dynamics, infrastructure features represent environmental constraints, and market indicators provide macro-level context. This structured integration preserves the heterogeneity of the original data while enabling joint modeling within a single predictive framework.

## 4 Methodology

### 4.1 Problem Formulation and CLV Definition

The objective of this study is to predict the customer lifetime value (CLV) of individual NEV users by integrating heterogeneous behavioral and contextual data sources. Let $i$ denote a customer and $t$ denote a discrete time period. CLV is defined as the discounted sum of expected future economic contributions generated by charging activities over a predefined horizon $H$:

$$CLV_i = \sum_{t=1}^{H} \frac{\mathbb{E}(R_{i,t}|X_i)}{(1+r)^t} \tag{1}$$

Where $R_{i,t}$ represents the revenue (or revenue proxy) generated by customer $i$ at time $t$, $r$ is the discount rate, and $\mathbf{X}_i$ denotes the set of multi-source features associated with the customer. This formulation follows standard CLV modeling practice while explicitly accounting for the usage-based value realization mechanism characteristic of the NEV charging ecosystem. In contrast to traditional retail contexts, where customer value is primarily realized through repeat purchases, NEV customer value emerges mainly from recurrent service usage. Charging behavior therefore provides a suitable and observable proxy for long-term economic contribution, allowing CLV prediction to be framed as a forward-looking regression problem under realistic data availability constraints. This formulation is consistent with standard CLV definitions while remaining flexible enough to accommodate the usage-based revenue structure of the NEV ecosystem. In contrast to traditional retail settings, NEV customer value is primarily realized through repeated charging transactions rather than discrete repurchase events. Therefore, the prediction task is framed as a regression problem in which future charging-derived revenue streams are estimated based on historical behavior and contextual information.

### 4.2 Feature Representation from Heterogeneous Data Sources

To capture the multidimensional nature of NEV customer behavior, features are constructed from three complementary data sources: charging transactions, charging infrastructure, and market context. Each source is processed independently to preserve its structural characteristics before fusion.Charging transaction features represent user-level behavioral dynamics. These include aggregated indicators such as charging frequency, total energy consumption, average session duration, and monetary expenditure within a given time window, as well as temporal descriptors capturing inter-session intervals and behavioral stability. Such features reflect both the intensity and regularity of customer engagement.Infrastructure features describe the charging environment faced by the user. Based on spatially aligned station data, indicators such as charging station density, proximity to fast-charging facilities, and infrastructure growth rates are computed at the regional level. These features capture supply-side constraints and opportunities that influence charging convenience and long-term usage patterns.Market context features represent macro-level adoption conditions, including regional NEV penetration rates and growth trends. These variables provide background information on market maturity and network effects, which may indirectly affect individual charging behavior and retention.By separating feature construction across data sources, the framework avoids premature aggregation and allows each modality to contribute distinct information to the prediction task.

### 4.3 Temporal and Spatial Alignment Strategy

Given the heterogeneous nature of the data, explicit temporal and spatial alignment is required before joint modeling. Temporal alignment is achieved by aggregating all features to a common time unit, ensuring consistency between historical inputs and future CLV targets. Spatial alignment is conducted by mapping charging sessions and infrastructure attributes to standardized regional units, which also serve as the basis for market context variables.This two-stage alignment strategy ensures that features from different sources are comparable and synchronized, reducing the risk of information leakage and spurious correlations. Importantly, alignment procedures rely solely on observable attributes such as timestamps and geographic identifiers, without introducing artificial smoothing or imputation beyond documented aggregation rules.

### 4.4 Modality-Aware Fusion Architecture

A central methodological challenge in this study lies in integrating heterogeneous data sources without collapsing their distinct informational roles. Rather than directly concatenating raw features from different sources, a modality-aware fusion architecture is adopted, in which each data modality is first transformed independently before integration. Let $\mathbf{X}_i^{(b)}$, $\mathbf{X}_i^{(s)}$ and $\mathbf{X}_i^{(m)}$ denote the behavioral, infrastructure, and market context feature sets for customer $i$ respectively. Each feature subset is first encoded independently through source-specific transformation functions:

$$\mathbf{h}_i^{(b)} = f_b(\mathbf{X}_i^{(b)}), \quad \mathbf{h}_i^{(s)} = f_s(\mathbf{X}_i^{(s)}), \quad \mathbf{h}_i^{(m)} = f_m(\mathbf{X}_i^{(m)}) \tag{2}$$

Where $f_b(\cdot), f_s(\cdot)$, and $f_m(\cdot)$ are encoding functions implemented using *feedforward* or sequence-based neural networks depending on feature structure. The encoded representations are then combined through a fusion layer that learns the relative contribution of each modality:

$$\mathbf{h}_i = g([\mathbf{h}_i^{(b)}, \mathbf{h}_i^{(s)}, \mathbf{h}_i^{(m)}]) \tag{3}$$

where $g(\cdot)$ denotes a nonlinear fusion function, and $[\cdot]$ represents vector concatenation. This architecture enables the model to learn cross-modal complementarities while maintaining a clear separation between heterogeneous information sources. Compared with conventional feature aggregation approaches, the proposed design explicitly accounts for modality heterogeneity and aligns behavioral and contextual information toward customer lifetime value prediction, thereby enhancing both interpretability and predictive robustness. This design allows the model to adaptively weight behavioral signals and contextual constraints, reflecting their joint influence on long-term value formation. Compared with simple feature concatenation, the modality-aware architecture explicitly accounts for heterogeneity in data origin and scale, aligning with established principles of structured data fusion.

### 4.5 Model Training and Prediction

The fused representation hi is used as input to a regression layer that estimates the expected future revenue stream or directly outputs the predicted CLV value. Model parameters are learned by minimizing a loss function based on prediction error, such as mean squared error, over the training dataset.To prevent overfitting and ensure generalization, regularization techniques and early stopping strategies are employed. Importantly, training and evaluation are conducted using time-based data splits, so that only historical information is used to predict future outcomes. This setting closely reflects real-world deployment scenarios and avoids optimistic bias in performance assessment.

## 5 Experimental Design and Evaluation

This section presents the experimental design and evaluation protocol adopted to assess the effectiveness of the proposed multi-source heterogeneous data fusion framework for customer lifetime value (CLV) prediction in the new energy vehicle (NEV) charging context. All experiments are conducted using publicly available charging session data, and the entire workflow—from data preprocessing to model evaluation—is designed to ensure transparency and reproducibility.

### 5.1 Data Preparation and Sample Construction

The empirical analysis is based on the ACN-Data charging session dataset, which records real-world electric vehicle charging behavior at multiple charging sites. After data cleaning procedures that remove incomplete sessions and records with missing key attributes, a total of 4,819 valid charging sessions associated with 366 anonymized users are retained. The data span the period from January 2020 to September 2025.Although the final supervised learning dataset consists of user–time forecasting instances, the analysis is grounded in several thousand charging sessions across hundreds of users. This multi-granularity structure allows charging behavior to be observed and summarized at the session level, the user level, and the user–time level, thereby providing a rich behavioral basis for CLV prediction.To formulate a supervised learning task consistent with the CLV definition, monthly anchor points are constructed for each user when sufficient historical observations are available. For each anchor time, charging behavior over the preceding 180 days is used to construct explanatory variables, while the target variable is defined as the discounted charging-based CLV over the subsequent 90 days. This procedure yields 150 user–time observations, which serve as the final analytical sample.

### 5.2 CLV Definition and Evaluation Setting

Customer lifetime value is operationalized as a usage-based revenue proxy derived from charging behavior. Specifically, CLV is computed as the discounted sum of charging revenue over a 90-day horizon, where revenue is approximated by delivered energy multiplied by a fixed per-unit electricity price. In addition to the primary 90-day forecasting horizon, two alternative horizons are considered to examine the robustness of the proposed framework: a short-term horizon of 30 days and a longer-term horizon of 180 days. For each horizon, CLV is defined consistently using the same revenue proxy and discounting scheme, with only the forecasting window adjusted accordingly.A monthly discount factor is applied to reflect the time value of future consumption. This definition relies exclusively

Table 1: CLV prediction performance

| Model | MAE | RMSE | Spearman $\rho$ |
|---|---|---|---|
| Model 1: Ridge (transaction aggregates) | 20.7302 | 28.7623 | 0.2765 |
| Model 2: Behavioral ML (ExtraTrees) | 19.9516 | 28.2708 | 0.2872 |
| Model 3: Fusion (ExtraTrees, proposed) | 19.0958 | 27.5130 | 0.3825 |

on observable variables provided in the public dataset and avoids the introduction of proprietary cost assumptions.To reflect real-world forecasting conditions and prevent information leakage, a time-based data split is employed. Observations with anchor times prior to January 2021 are used for model training, while later observations form the test set. This chronological separation ensures that all predictions are generated using strictly historical information.

### 5.3 Model Configuration and Baselines

Three models are implemented for comparison. The first model represents a traditional CLV regression based on aggregated transaction indicators, serving as a benchmark for transaction-centric approaches. The second model applies a machine learning regressor to a richer set of behavioral features derived from charging sessions, capturing nonlinear usage patterns. The third model corresponds to the proposed fusion framework, which augments behavioral features with station- and site-level interaction descriptors that characterize users' spatial charging diversity.All models are trained under identical data splits and evaluation criteria. Hyperparameters are selected using validation performance within the training period, and no information from the test period is used during model tuning.

### 5.4 Evaluation Metrics

Model performance is assessed using a combination of accuracy- and ranking-oriented metrics. Mean absolute error (MAE) and root mean squared error (RMSE) quantify numerical prediction accuracy, while Spearman's rank correlation coefficient evaluates the consistency between predicted and realized CLV rankings. This combination reflects both statistical accuracy and practical usefulness in value-based customer prioritization.

## 6 Results and Discussion

### 6.1 Overall Predictive Performance

Table 1 reports the predictive performance of the three models under the primary 90-day forecasting horizon. Customer lifetime value is defined as the discounted charging-based revenue over the next 90 days, where revenue is approximated by the amount of electricity charged multiplied by a fixed price of 0.20 USD/kWh and discounted at a monthly rate of 1%. For each user, features are constructed from charging behavior observed during the preceding 180 days, using the beginning of each month as the anchor point. A time-based split is adopted, with observations anchored before 2021-01-01 used for training and later observations reserved for testing. The results reveal a clear performance hierarchy across the models. The fusion-based model achieves the lowest MAE and RMSE, together with a substantially higher Spearman's rank correlation, compared with both baseline approaches. In particular, the improvement in rank correlation indicates that integrating heterogeneous information enhances the model's ability to correctly order customers by long-term value, which is critical for value-based decision support. These findings suggest that CLV in the NEV charging context is not solely determined by physical consumption intensity, but instead reflects a combination of behavioral regularities and contextual interaction patterns that are explicitly captured by the proposed fusion framework.

### 6.2 Prediction Accuracy Visualization

Figure 1 visualizes the relationship between predicted and realized CLV values produced by the fusion model under the 90-day forecasting horizon. Most observations cluster around the diagonal

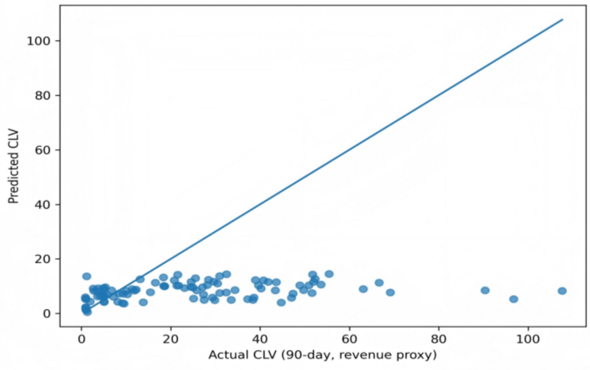


Figure 1: Predicted vs Actual CLV (Fusion Model)

reference line, indicating a strong correspondence between predicted and observed values. Larger deviations are primarily associated with users exhibiting sparse or irregular charging histories, highlighting the inherent uncertainty in forecasting future behavior when historical information is limited.

### 6.3 Robustness Across Forecasting Horizons

To examine the robustness of the proposed framework under different decision time scales, model performance is evaluated across three forecasting horizons: 30, 90, and 180 days. For all horizons, the historical feature window is fixed at 180 days, and the same time-based train–test split is applied to ensure strict comparability across settings.

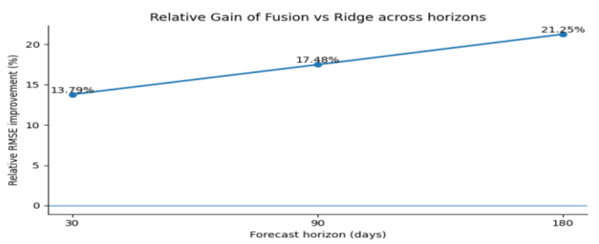


Figure 2: Robustness of CLV Prediction Performance Across Forecasting Horizons

Figure 2 summarizes how prediction performance evolves as the forecasting horizon extends. As expected, numerical prediction accuracy deteriorates monotonically with increasing horizon length for all models, reflecting the accumulation of uncertainty inherent in long-term usage-based value forecasting. This pattern is consistent with prior findings in CLV and demand forecasting literature, where longer horizons amplify behavioral volatility and data sparsity effects.Despite this overall degradation trend, clear differences emerge across modeling approaches. The fusion-based model consistently exhibits lower RMSE than transaction-centric and behavioral-only baselines at all horizons. More importantly, the rate of performance deterioration differs substantially across models. While baseline models experience a steep increase in prediction error as the horizon extends from 30 to 180 days, the fusion model shows a noticeably flatter error growth curve, indicating greater robustness to horizon expansion.

The relative gain analysis in Figure 2 further clarifies this robustness advantage. Compared with baseline models, heterogeneous data fusion yields modest improvements at the short-term horizon (30 days), where recent transaction intensity already provides strong predictive signal. The relative improvement becomes most pronounced at the medium-term horizon (90 days), suggesting that infrastructure interaction features and behavioral regularity descriptors play a critical role in stabilizing value estimation beyond immediate consumption patterns. At the longest horizon (180 days), although absolute prediction errors increase for all models, the fusion framework continues to deliver positive relative gains, demonstrating its ability to retain non-trivial predictive signal in sparse and uncertain long-term settings.

These results highlight an important distinction between short-, medium-, and long-term CLV prediction tasks. Short-horizon forecasts primarily benefit from recent transactional information, whereas medium-horizon predictions gain substantially from the integration of heterogeneous behavioral and contextual features. Long-horizon forecasts remain inherently challenging due to irregular usage patterns, but the proposed fusion framework mitigates error growth more effectively than baseline approaches. Overall, the horizon-based evaluation confirms that heterogeneous data fusion improves not only point accuracy but also robustness across decision-relevant forecasting horizons.

### 6.4 Contribution of Heterogeneous Features

To further understand how different information sources contribute to customer lifetime value (CLV) prediction, an ablation analysis is conducted by selectively removing heterogeneous feature groups from the fusion framework. The analysis focuses on infrastructure interaction features derived from charging session metadata, while keeping the modeling architecture, training procedure, and evaluation protocol unchanged. This controlled setting ensures that observed performance differences can be attributed to the presence or absence of specific feature groups rather than to model configuration effects.

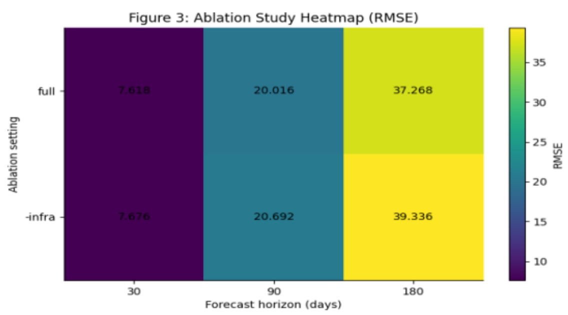


Figure 3: Ablation Heatmap of Heterogeneous Feature Contributions Across Forecasting Horizons

Figure 3 presents the ablation results in the form of a heatmap across multiple forecasting horizons. Each cell reports the increase in RMSE relative to the full fusion model when a particular feature group is excluded. Across all horizons, removing infrastructure interaction features consistently leads to a noticeable degradation in prediction accuracy, confirming that these features provide meaningful explanatory power beyond individual charging behavior alone. The magnitude of error increase varies with the forecasting horizon, indicating that the contribution of heterogeneous features is horizon-dependent rather than uniform over time.The impact of infrastructure-related features becomes more pronounced as the forecasting horizon extends from short-term to medium-term settings. At the 30-day horizon, performance degradation remains limited, suggesting that recent charging intensity and behavioral aggregates capture much of the short-term predictive signal. At the 90-day horizon, however, the exclusion of infrastructure interaction features results in a substantially larger increase in RMSE, highlighting the importance of spatial charging diversity and infrastructure accessibility in stabilizing medium-term value estimation. This pattern aligns with the operational characteristics of the NEV charging ecosystem, where sustained usage depends not only on past consumption but also on the feasibility of continued interaction with charging facilities.At the longest horizon (180 days), prediction errors increase for all model variants, reflecting the inherent uncertainty of long-term usage-based forecasting. Nevertheless, the relative degradation observed in the ablation setting remains significant, indicating that infrastructure interaction features continue to contribute non-trivial information even under sparse and irregular future behavior. These findings suggest that heterogeneous features do not merely refine short-term predictions, but instead help mitigate error growth as uncertainty accumulates over time.

Overall, the ablation results confirm that heterogeneous information sources contribute complementary rather than redundant predictive signals. Behavioral features capture individual usage intensity and regularity, while infrastructure interaction descriptors encode structural constraints and opportunities that shape long-term charging behavior. By integrating these heterogeneous feature groups within a unified fusion framework, the proposed approach achieves more robust CLV prediction across multiple forecasting horizons.

### 6.5 Value-Based Ranking and Lift Analysis

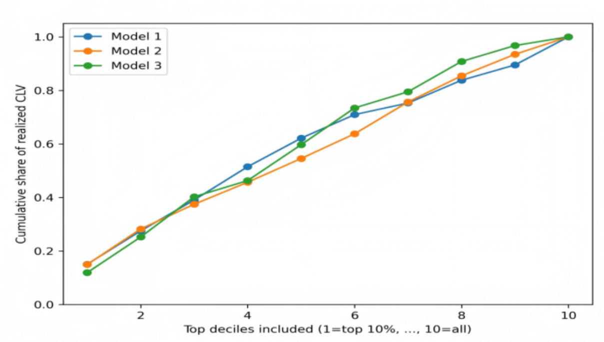


Figure 4: Decile-based Lift (Cumulative)

Figure 4 presents a decile-based cumulative lift analysis based on predicted CLV rankings generated by the fusion model. Users ranked in the top deciles account for a disproportionately large share of realized CLV, demonstrating that the model effectively concentrates value among a small subset of customers. This result highlights the practical relevance of the proposed framework for value-based segmentation tasks, such as identifying high-priority users for targeted services, pricing strategies, or infrastructure planning.

### 6.6 Discussion

This study examines the role of heterogeneous behavioral and contextual information in customer lifetime value (CLV) prediction within the new energy vehicle (NEV) charging context. By integrating charging behavior with infrastructure interaction descriptors derived from publicly available data, the proposed framework provides a structured approach to usage-based CLV estimation. The following discussion focuses on interpreting the observed performance patterns, clarifying the mechanisms underlying the fusion effects, and outlining the boundaries of applicability.

A first key insight is that CLV in the NEV charging ecosystem cannot be adequately characterized using transaction aggregates alone. While historical charging frequency and energy consumption remain informative, their explanatory power weakens as the evaluation horizon extends. This observation is consistent with established findings in CLV research that highlight the limitations of purely transactional representations in non-contractual and usage-based settings [1]-[3]. In the charging context, long-term value realization depends not only on accumulated consumption, but also on the regularity and feasibility of continued usage. Infrastructure interaction patterns therefore capture structural constraints and opportunities that shape future behavior, complementing individual usage histories.

The horizon-based analysis further reveals a clear distinction between absolute value prediction and value-based ranking. Prediction errors increase monotonically as the forecasting horizon extends from 30 to 180 days, reflecting the accumulation of uncertainty inherent in long-term behavioral forecasting. Such horizon sensitivity has been widely documented in demand forecasting and CLV modeling, where longer prediction windows amplify the effects of behavioral volatility [4]. At the same time, ranking consistency exhibits a non-monotonic pattern, with the most reliable customer ordering observed at intermediate horizons. This suggests that while heterogeneous data fusion improves numerical accuracy over extended horizons, stable value-based prioritization is most feasible within medium-term decision windows, which are typically most relevant for operational decision support.

The ablation results shed light on the mechanisms through which heterogeneous data fusion improves performance. Infrastructure-related features, including spatial diversity and interaction intensity with charging facilities, contribute more substantially to error reduction than macro-level contextual indicators. This finding reflects the operational reality of NEV charging systems, where accessibility and spatial flexibility directly influence users' ability to sustain regular charging behavior. Similar relationships between infrastructure availability and usage persistence have been observed in energy and mobility systems, underscoring the importance of incorporating supply-side interaction patterns into value modeling frameworks [6, 7]. Market-level context variables, while less influential, provide complementary background information that conditions demand without fully determining individual trajectories.An important methodological implication of these findings lies in the stability of feature contributions across time. The consistency of feature importance rankings between training and test periods suggests that the fusion model captures persistent behavioral and infrastructural relationships rather than transient correlations. This stability is particularly relevant when working with public datasets, which often exhibit sparsity and irregular observation patterns. From a data fusion perspective, the results support modality-aware integration strategies that preserve the distinct roles of heterogeneous information sources, rather than collapsing them into undifferentiated feature sets[8]-[10].

Several limitations of the present study should be acknowledged. First, CLV is operationalized using a charging-based revenue proxy with fixed pricing and discounting assumptions. Although this approach ensures transparency and reproducibility and is consistent with prior usage-based CLV studies [11]-[13], it does not account for pricing heterogeneity, cost structures, or cross-service interactions that may affect net profitability. Second, the empirical analysis is based on a single charging network and a specific temporal window. Behavioral responses to infrastructure conditions may differ across regions with alternative regulatory frameworks or charging technologies. Third, the decline in ranking stability at long horizons suggests that CLV-based prioritization should be complemented with adaptive or rolling evaluation strategies in practice.

These limitations point to several avenues for future research. Extending the framework to multi-network datasets, incorporating dynamic pricing information, or integrating real-time infrastructure utilization metrics may further enhance predictive robustness and external validity. In addition, combining heterogeneous data fusion with sequential or survival-based modeling approaches may improve long-horizon ranking stability while maintaining interpretability. Importantly, such extensions can be pursued without sacrificing transparency, which remains essential for cumulative research in data-driven decision support.

Overall, this study demonstrates that heterogeneous data fusion based on publicly available charging and infrastructure data provides a viable and methodologically grounded approach to CLV prediction in the NEV domain. Rather than introducing algorithmic complexity for its own sake, the proposed framework illustrates how structured integration of behavioral and contextual information can yield robust improvements in predictive accuracy and practical relevance, contributing to both CLV research and the broader literature on data fusion and decision-support systems.

# 7 Conclusion and Future Research

This study investigated customer lifetime value (CLV) prediction in the new energy vehicle (NEV) charging context by integrating heterogeneous behavioral and contextual information derived exclusively from publicly available data. Motivated by the limitations of transaction-based CLV models in usage-driven and non-contractual environments, a modality-aware data fusion framework was proposed to combine charging behavior, infrastructure interaction, and market context features within a unified predictive setting.Empirical results demonstrate that the proposed framework yields consistent improvements over conventional baselines in terms of absolute prediction accuracy and medium-horizon ranking performance. In particular, the horizon-based evaluation shows that heterogeneous data fusion enhances robustness across multiple forecasting windows, while ablation and feature stability analyses clarify the complementary roles of behavioral regularity and infrastructure interaction patterns. These findings confirm that long-term customer value in the NEV charging ecosystem is shaped not only by historical consumption intensity, but also by the structural conditions under which charging behavior can be sustained.From a methodological perspective, this study contributes to the CLV and data fusion literature by illustrating how heterogeneous information sources can be integrated without resorting to excessive model complexity. Rather than relying on deep or opaque architectures, the proposed approach emphasizes structured feature construction, modality-aware representation, and transparent evaluation. This design choice aligns with the requirements of reproducibility and interpretability, particularly when working with public datasets in decision-support contexts.The study also offers practical implications for stakeholders involved in the planning and management of charging infrastructure and customer services. The results suggest that value-based customer assessment should account for infrastructure accessibility and interaction diversity, especially when decisions are made over medium-term horizons. At the same time, the observed decline in ranking stability for long-horizon predictions highlights the need for adaptive evaluation strategies when CLV estimates are used for operational prioritization.Several limitations provide directions for future research. First, CLV was operationalized using a charging-based revenue proxy with fixed pricing and discounting assumptions. Incorporating dynamic pricing schemes, cost heterogeneity, or cross-service interactions would allow a more comprehensive assessment of customer profitability. Second, the empirical analysis was based on a single charging network and a specific temporal window. Extending the framework to multi-network or cross-regional datasets would help evaluate its generalizability under different regulatory and infrastructural conditions. Third, while the current framework focuses on static aggregation within fixed windows, future work could explore sequential or survival-based modeling approaches to further improve long-horizon prediction stability.

In conclusion, this study demonstrates that heterogeneous data fusion using transparent and reproducible methods provides a viable pathway for advancing CLV prediction in emerging energy and mobility systems. By balancing methodological rigor with practical relevance, the proposed framework contributes to both academic research and data-driven decision support in the evolving NEV ecosystem.

## References


[1] Fader, P.S.; Hardie, B.G.S.; Lee, K.L. (2005). RFM and CLV: Using iso-value curves for customer base analysis, *Journal of Marketing Research*, 42(4), 415–430, 2005.

[2] Schmittlein, D.C.; Morrison, D.G.; Colombo, R. (1987). Counting your customers: Who are they and what will they do next?, *Management Science*, 33(1), 1–24, 1987.

[3] Fader, P.S.; Hardie, B.G.S.; Shang, J. (2010). Customer-base analysis in a discrete-time noncontractual setting, *Marketing Science*, 29(6), 1086–1108, 2010.

[4] Rossi, F.; Basso, L.; Cattaneo, C. (2019). Machine learning approaches for customer lifetime value prediction, *Expert Systems with Applications*, 129, 156–170, 2019.

[5] Bertsimas, D.; Kallus, N. (2020). From predictive to prescriptive analytics, *Management Science*, 66(3), 1025–1044, 2020.

[6] Shmueli, G.; Koppius, O.R. (2011). Predictive analytics in information systems research, *MIS Quarterly*, 35(3), 553–572, 2011.

[7] Neslin, S.A.; Gupta, S.; Kamakura, W.; Lu, J.; Mason, C.H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models, *Journal of Marketing Research*, 43(2), 204–211, 2006.

[8] Kumar, V.; Reinartz, W. (2018). *Customer Relationship Management: Concept, Strategy, and Tools*, Springer, Berlin, Germany, 2018.

[9] Sun, M.; Li, S. (2021). Context-aware consumer behavior modeling with heterogeneous data, *Decision Support Systems*, 142, Article 113473, 2021.

[10] Hall, D.L.; Llinas, J. (1997). An introduction to multisensor data fusion, *Proceedings of the IEEE*, 85(1), 6–23, 1997.

[11] Chen, J.I.-Z.; Chung, Y.-N. (2012). A data fusion methodology for wireless sensor systems, *International Journal of Computers Communications & Control*, 7(1), 39–52, 2012. DOI: 10.15837/ijccc.2012.1.208

[12] Zhang, M.; Lu, J.; Ma, N.; Cheng, T.C.E.; Hua, G. (2022). A feature engineering and ensemble learning based approach for repeated buyers prediction, *International Journal of Computers Communications & Control*, 17(6), Article 4988, 2022. DOI: 10.15837/ijccc.2022.6.4988

[13] Hu, F. (2025). An approach of sentiment-topic mining based on user online comments on new energy vehicles, *International Journal of Computers Communications & Control*, 20(6), Article 6934, 2025. DOI: 10.15837/ijccc.2025.6.6934

[14] Shmueli, G.; Bruce, P.; Yahav, I.; Patel, N.R.; Lichtendahl, K.C. (2020). *Data Mining for Business Analytics*, Wiley, Hoboken, NJ, USA, 2020.

[15] Burez, J.; Van den Poel, D. (2009). Handling class imbalance in customer churn prediction, *Expert Systems with Applications*, 36(3), 4626–4636, 2009. DOI: 10.1016/j.eswa.2008.05.027

[16] Verbeke, W.; Dejaeger, K.; Martens, D.; Hur, J.; Baesens, B. (2011). Predictive modeling for customer churn, *Decision Support Systems*, 52(1), 164–173, 2011. DOI: 10.1016/j.dss.2011.05.011

[17] Sadeghianpourhamami, N.; Refa, N.; Develder, C.; Strobbe, M. (2019). Electric vehicle charging demand forecasting using real-world data, *Applied Energy*, 238, 1507–1518, 2019. DOI: 10.1016/j.apenergy.2019.01.083

[18] Liu, Z.; Wen, F.; Ledwich, G. (2013). Optimal planning of electric-vehicle charging stations, *IEEE Transactions on Power Delivery*, 28(1), 354–362, 2013.

[19] Shao, C.; Yang, H.; Zhang, Y.; Ke, J. (2013). A simple approximation for the charging demand of electric vehicles, *Transportation Research Part D: Transport and Environment*, 21, 7–13, 2013. DOI: 10.1016/j.trd.2013.01.004

[20] Li, X.; Zhang, Y. (2020). Mining user perceptions of electric vehicles from online reviews, *Energy Policy*, 138, Article 111245, 2020.

[21] Lee, Z.J.; Li, T.; Low, S.H. (2019). ACN-Data: Analysis and applications of an open electric vehicle charging dataset, *Energy Informatics*, 2(1), Article 16, 2019. DOI: 10.1186/s42162-019-0079-9

[22] National Renewable Energy Laboratory (NREL). (2023). Alternative Fueling Station Locations, U.S. Department of Energy, Open Dataset, 2023.

[23] Gupta, S.; Lehmann, D.R.; Stuart, J.A. (2004). Valuing customers, *Journal of Marketing Research*, 41(1), 7–18, 2004. DOI: 10.1509/jmkr.41.1.7.25084

[24] Batislam, E.P.; Denizel, M.; Filiztekin, A. (2007). Empirical validation and comparison of models for customer base analysis, *International Journal of Research in Marketing*, 24(3), 201–209, 2007. DOI: 10.1016/j.ijresmar.2006.12.005

[25] Hardman, S.; et al. (2018). A review of consumer preferences of and interactions with electric vehicle charging infrastructure, *Transportation Research Part D: Transport and Environment*, 62, 508–523, 2018. DOI: 10.1016/j.trd.2018.04.002

[26] Nicholas, M.A.; Tal, G.; Turrentine, T. (2017). Advanced plug-in electric vehicle travel and charging behavior interim report, *Transportation Research Record*, 2654(1), 1–10, 2017. DOI: 10.3141/2654-01

[27] Hyndman, R.J.; Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd ed.), OTexts, 2021.

[28] Athey, S. (2018). The impact of machine learning on economics, In: *The Economics of Artificial Intelligence*, University of Chicago Press, 2018.



Member since 2012
JM08090

This journal is a member of, and subscribes to the principles of,
the Committee on Publication Ethics (COPE).
https://publicationethics.org/members/international-journal-computers-communications-and-control

*Cite this paper as:*

Mei, L.; Shang, Y.; Deng, Y.; Fan, L.; Zhang, Z.; Yang, Y. (2026). Multi-Source Heterogeneous Data Fusion for Predicting Customer Lifetime Value in the New Energy Vehicle Market, *International Journal of Computers Communications* & *Control*, 21(5), 7431, 2026.
https://doi.org/10.15837/ijccc.2026.5.7431