International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 07 | July - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Vehicle Count Forecasting: A Gradient Boosting Approach on Simulated Data

Bindheyashrita Pradhan
Email Id: bpradhan2023@gift.edu.in

Prof. Smruti Smaraki Sarangi
Email Id: smrutismaraki@gift.edu.in

Abstract*:* Accurate urban vehicle count forecasting is essential for effective Intelligent Transportation Systems (ITS). Traditional models like ARIMA and Prophet struggle with complex urban traffic due to real-world data's 'complicated raw data' characteristics: high-dimensionality, non-linear dependencies, and intricate influences from external factors (festivals, weather, accidents). The challenge is teaching models to discern patterns in such dynamic, opaque environments. This study introduces a novel, integrated forecasting framework. Firstly, a sophisticated synthetic data generator was developed to explicitly produce 'complicated raw data.' It simulates hourly traffic for Bhubaneswar (2023–2024), precisely modeling time-varying multiplicative impacts of festivals, weather, zones, and accidents across vehicle types. This controlled complexity ensures rigorous model training. Secondly, a powerful LightGBM model is employed with advanced feature engineering. It uniquely integrates explicit event multiplier features, derived directly from the generator's logic, and key interaction terms. These enable the model to learn nuanced traffic behaviors and accurately forecast event-driven surge magnitudes. Evaluated on a 2023/2024 split, LightGBM (MAE-optimized) delivered high accuracy, outperforming Prophet significantly (e.g., Two-Wheelers MAE: 101.44 vs. Prophet's 623.40). A robust recursive strategy generated granular 2025 predictions. This demonstrates a superior pathway to overcome data limitations and achieve robust urban traffic forecasts.

Keywords: Vehicle Count Forecasting, Gradient Boosting, LightGBM, Synthetic Data Generation, Feature Engineering, Intelligent Transportation Systems (ITS), Multiplicative Event Impacts

I. INTRODUCTION

Rapid urban population growth has intensified traffic congestion, imposing economic, environmental, and social burdens. Extended travel times, higher fuel costs, and increased emissions undermine productivity and quality of life. Addressing these challenges requires intelligent traffic management, where accurate predictions are critical for proactive decision-making in Intelligent Transportation Systems (ITS). Traffic forecasting enables dynamic signal adjustments, optimized logistics, and real-time navigation, improving efficiency and reducing environmental impacts. While early methods relied on historical averages or basic statistical models, modern urban traffic's complexity demands advanced, data-driven approaches. This paper introduces an end-to-end framework for urban vehicle count forecasting, combining synthetic data generation and feature engineering with a Gradient Boosting model (LightGBM) to deliver high-accuracy predictions in complex urban environments.

II. LITERATURE REVIEW

Accurate urban vehicle count forecasting is crucial for Intelligent Transportation Systems (ITS), yet complex traffic dynamics, non-linear dependencies, and real-world data limitations pose significant challenges. Historically, classical statistical models like AutoRegressive Integrated Moving Average (ARIMA) [7] and Exponential Smoothing, alongside traditional machine learning techniques such as Support Vector Machines (SVR) [8], were employed. While capable of capturing linear temporal patterns and basic seasonality, these methods often struggled with the intricate, non-linear interactions of diverse external factors on traffic flow.

More advanced Machine Learning and Deep Learning (DL) architectures, including Recurrent Neural Networks like Long Short-Term Memory (LSTM) [5] and spatio-temporal networks combining Convolutional Neural Networks (CNNs) and Transformers [1, 2, 9], have emerged to capture complex dependencies and integrate contextual information. Studies have incorporated factors such as rainfall [3], general weather impacts [10], and planned events [1, 2]. However, these sophisticated models often demand vast, clean real-world datasets which are frequently scarce and incomplete—and can be computationally intensive and lack interpretability.

Gradient Boosting Machines (GBMs), particularly LightGBM [6], offer a powerful and efficient alternative, balancing high predictive accuracy with faster training times for structured datasets [13]. Despite progress, a crucial, less explored area is effectively leveraging precise, simulation-defined multiplicative impacts of events in feature engineering, beyond simple binary flags. Synthetic data generation [4, 11, 12] provides a vital controlled environment to overcome real-world data scarcity and validate models against complex, known dynamics. This research addresses this gap by explicitly integrating simulation knowledge into LightGBM's feature engineering to accurately predict event-driven traffic surges.

III. RESEARCH GAPS / PROBLEM FORMULATION

Accurate urban vehicle count forecasting is crucial for Intelligent Transportation Systems but inherently challenging. Urban traffic exhibits high intrinsic complexity, characterized by non-linear dynamics and intertwined influences from diverse external factors (weather, accidents, and particularly, major events that can have multiplicative impacts).

A primary impediment is the pervasive scarcity and heterogeneity of real-world traffic data, often lacking comprehensive detail and explicit ground truth on the

magnitude and time-varying profile of these external influences. Existing forecasting models, especially simpler additive ones like Prophet, frequently underpredict the scale of significant, event-driven traffic surges because they cannot explicitly leverage such impact magnitudes. This poses a key research gap: how to effectively train advanced models to accurately capture these complex, high-magnitude event-driven surges when explicit real-world impact data is limited. We formulate this problem by proposing a methodology that utilizes a sophisticated synthetic data generator, explicitly modeling multiplicative event effects. We then develop a novel feature engineering strategy that directly leverages this simulation's underlying logic to derive quantitative event multipliers, enabling a LightGBM model to accurately predict the scale and timing of these complex traffic patterns.

IV. PROPOSED METHODOLOGY / SYSTEM DESIGN

This research introduces a complete, end-to-end framework to forecast urban vehicle counts, as illustrated in the system architecture diagram (Figure 1). The methodology is designed as a sequential pipeline that transforms raw simulated inputs into a final, usable forecast.

The process begins with the Synthetic Data Generation Module (1), which uses a custom Python script to produce a raw, multi-factor dataset simulating hourly traffic with complex contextual factors. These include varied locations (e.g., Master Canteen, Unit 1, each with distinct base traffic patterns), diverse vehicle types (e.g., two-wheelers, three-wheelers, four-wheelers, and six-wheelers, each with unique volume profiles), assigned functional zones (e.g., Market Zones, Residential Areas, influencing traffic behavior), distinct weather conditions (e.g., sunny, rainy, foggy, impacting flow), and explicit details on simulated accidents (e.g., represented by a binary flag and summary string). Crucially, it models major events (e.g., festivals like Durga Puja, Rath Yatra, Diwali) with predefined multiplicative impacts, causing traffic to surge significantly (e.g., 3x or 4x increase) during specific hours. This raw data is then fed into the **Data Preparation & Feature Engineering Module (2)**. Here, the data is cleaned, aggregated, and transformed into a feature-rich matrix. Crucial features—such as temporal patterns (hour, day), lagged values, rolling statistics, and event multipliers—are engineered. The target vehicle count is also log-transformed to stabilize variance and prepare it for modeling.

The resulting data is split for training and validation. The training data is used by the **LightGBM Model Training Module (3),** which is optimized to minimize Mean Absolute Error (MAE). This module learns the complex, non-linear relationships within the data, producing the essential Trained Model artifact. The model's predictive power is verified against a test set using the **Evaluation Module (4).**
For generating future predictions, the **Recursive Forecasting Module (5)** utilizes the Trained Model. It operates iteratively, predicting one future time step at a time. For each new prediction, it generates the required input features by combining simulated future context (e.g., holidays) with dynamic features (e.g., lagged values) derived from its own previously generated forecasts.

Finally, the stream of log-scale predictions is processed by the **Output & Visualization Module (6).** It performs an inverse transformation to convert predictions back to their original vehicle count scale and formats the results into a final, human-readable **CSV output**, completing the workflow.

Key innovations include:
Systematic translation of simulation logic into predictive features, Explicit modeling of multiplicative event impacts, Robust handling of complex, non-linear interactions through gradient boosting
This methodology bridges the gap between traffic simulation knowledge and data-driven forecasting, particularly for challenging event-driven scenarios.
The framework was implemented in Python using LightGBM (4.6.0), Pandas, and NumPy. A synthetic dataset simulated Bhubaneswar traffic (2023-2024) with: Hourly vehicle counts (4 categories), Multiplicative event impacts (festivals, accidents), Weather/zone influences

- Key steps: Generated 17,520 hourly records with embedded complexities, Engineered temporal, lagged, and event-multiplier features, Trained LightGBM (MAE-optimized) on 2023 data, tested on 2024 ,Generated 2025 forecasts via recursive prediction **- Hardware:** Standard PC (Ryzen 3, 8GB RAM).

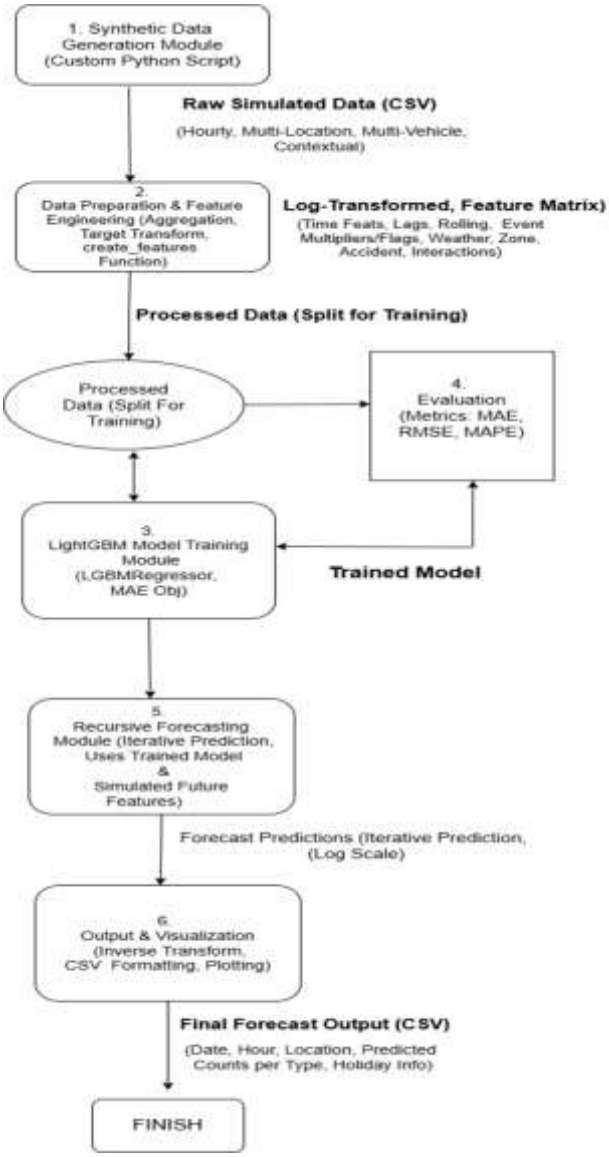

Figure 1: System Architecture

V. RESULTS AND DISCUSSION

The core comparison focuses on the MAE performance of LightGBM versus the baseline Prophet model. **Table 1** directly highlights this. Prophet, while capturing general seasonality, significantly underestimated high-impact event spikes on a 2024-10-01 onwards test set.

Vehicle Type	Prophet MAE	LightGBM MAE
Two-Wheelers	623.40	101.44
Three-Wheelers	2392.92	395.43
Four-Wheelers	326.22	51.29
Six-Wheelers	270.09	46.10

Table 1 : MAE Comparison Between Prophet and LightGBM (2024 Test Period)

Visual Analysis and Forecast Characteristics: Visualizations provide essential qualitative insights into the models' performance, pattern capture capabilities, and the nature of the generated forecasts over time. Due to space constraints, selected representative plots for Two-Wheelers are presented here; comprehensive visualizations for all vehicle types are available in the full thesis document. **Figure 2** displays the comparison between actual Two-Wheelers counts and the baseline FB Prophet model's predictions over the 2024-10-01 onwards test period. This plot clearly shows the significant underestimation by Prophet for Two-Wheelers during major event peaks in October 2024, highlighting its limitations with multiplicative effects.

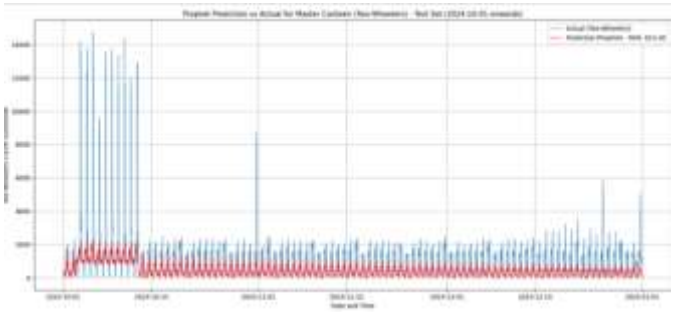

Figure 2: Prophet Prediction vs Actual for Two-Wheelers at Master Canteen (2024-10-01)

In stark contrast, **Figure 3** demonstrates LightGBM's superior visual fit for Two-Wheelers over the full 2024 test set. This graph accurately captures both the timing and magnitude of major event surges throughout the year (e.g., Rath Yatra in July, Durga Puja in October, Diwali in November, Christmas/NYE in December), indicating the model's robustness and enhanced feature engineering effectiveness.

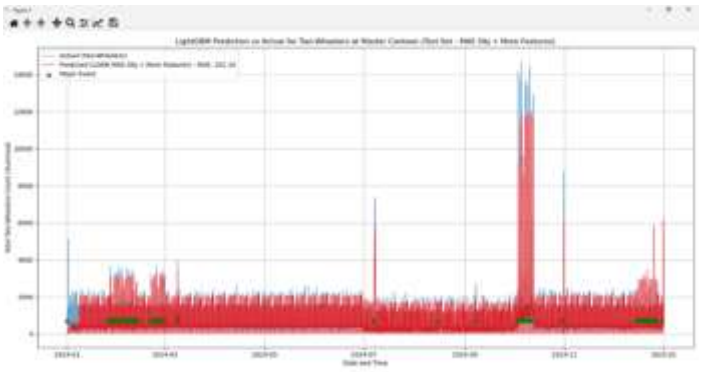

Figure 3: LightGBM Prediction vs Actual for Two-Wheelers at Master Canteen (Full 2024 Test Set - MAE Obj + Enhanced Features)

Finally, **Figure 4** illustrates the hourly Two-Wheelers forecast for the entire year 2025, generated by the recursive forecasting methodology. This plot displays clear daily and weekly seasonality, consistent with learned patterns, and crucially predicts substantial traffic surges during periods corresponding to known major events in 2025, with predicted magnitudes aligning well with the event impacts learned from historical data. This forecast demonstrates a plausible projection of traffic patterns into the future.

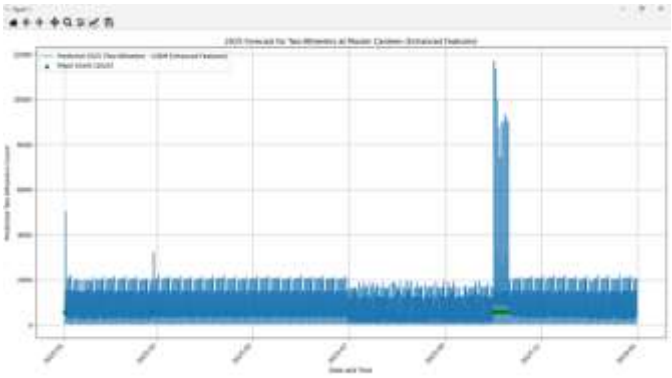

Figure 4: 2025 Forecast for Two-Wheelers at Master Canteen (LightGBM Enhanced Features)

2025 Forecast Results (Single Location) The recursive LightGBM model generated hourly vehicle count predictions for Master Canteen (2025). Example CSV structure:

	A	B	C	D	E	F	G	H
1	Date	Hour	Location	Two-Whee	Three-Wh	Four-Whe	Six-Wheel	Holiday_Name
2	01-01-2025	0	Master Ca	928.7379	4028.071	510.2202	321.1339	New Year
3	01-01-2025	1	Master Ca	1004.06	3851.8	613.864	277.6707	New Year
4	01-01-2025	2	Master Ca	796.5962	2874.509	449.6853	308.1505	New Year
5	01-01-2025	3	Master Ca	704.5091	2836.447	432.2703	262.6147	New Year
6	01-01-2025	4	Master Ca	670.6088	2470.064	394.1809	216.7593	New Year
7	01-01-2025	5	Master Ca	815.279	2877.85	442.2182	221.3376	New Year
8	01-01-2025	6	Master Ca	840.3191	2933.237	400.9014	258.6859	New Year
9	01-01-2025	7	Master Ca	54.46716	225.1425	25.12132	23.68901	New Year
10	01-01-2025	8	Master Ca	57.89286	227.9278	21.89875	25.99844	New Year
11	01-01-2025	9	Master Ca	56.3402	237.6749	27.7937	25.63456	New Year
12	01-01-2025	10	Master Ca	1048.259	5963.791	747.7965	737.599	New Year
13	01-01-2025	11	Master Ca	1508.078	5993.813	748.6256	743.3683	New Year
14	01-01-2025	12	Master Ca	1487.608	5993.292	751.7057	725.6225	New Year
15	01-01-2025	13	Master Ca	1546.333	5923.565	759.7579	537.2361	New Year
16	01-01-2025	14	Master Ca	1486.89	5949.088	761.0417	766.2327	New Year
17	01-01-2025	15	Master Ca	1516.427	5908.419	757.7639	749.1672	New Year
18	01-01-2025	16	Master Ca	585.7837	2143.656	296.3631	202.6266	New Year
19	01-01-2025	17	Master Ca	4963.082	17600.1	2886.296	1794.466	New Year
20	01-01-2025	18	Master Ca	5020.951	17770.42	2204.729	1848.432	New Year
21	01-01-2025	19	Master Ca	4919.655	17623.26	2925.082	1821.314	New Year
22	01-01-2025	20	Master Ca	1064.814	4305.795	666.5123	376.7544	New Year
23	01-01-2025	21	Master Ca	1196.538	5180.269	798.2116	498.6138	New Year
24	01-01-2025	22	Master Ca	1254.429	5148.685	786.672	458.6837	New Year
25	01-01-2025	23	Master Ca	1211.123	4226.954	723.7679	460.4367	New Year
26	02-01-2025	0	Master Ca	455.7264	1861.515	193.4721	217.685	None
27	02-01-2025	1	Master Ca	388.406	1642.723	210.1361	196.7132	None
28	02-01-2025	2	Master Ca	401.2719	1585.937	169.4672	172.3625	None

Figure 5: 2025 forecast CSV file containing all vehicle types (two-wheelers, three-wheelers, four-wheelers, and six-wheelers) for a single location.

VI. DISCUSSION OF FINDINGS

The superior predictive power of LightGBM stems directly from the comprehensive and domain-informed feature engineering strategy. While standard temporal and lagged features formed a strong base, the Event Multiplier features, derived explicitly from the generator's rules, were pivotal. These features enabled the model to learn and scale predictions proportionally to event magnitudes, effectively capturing the multiplicative impacts that additive models like Prophet cannot. Furthermore, interaction features successfully modeled how different factors combine non-linearly, such as the synergistic effect of a festival occurring during peak hours.

Configuring the model with an MAE objective (regression_l1) was also a critical strategic choice. This metric is inherently less sensitive to the quadratic penalization of large errors seen in RMSE, making it more robust for volatile, event-driven data and guiding the model to a practically interpretable solution. The strong performance on the 2024 test set after training only on 2023 data serves as a powerful validation of the synthetic data approach itself, confirming that the generated data effectively captures the underlying dynamics needed for advanced model development and offering a viable template for overcoming real-world data scarcity.

While visual results for the Two-Wheeler category are presented as a representative illustration, the model's success extends across all vehicle types, as confirmed by the strong quantitative metrics in the results table. The final 2025 forecast, generating simultaneous predictions for all categories as shown in the output **CSV file (Figure 5),** is the ultimate proof of the methodology's comprehensive viability. The model successfully captures the unique seasonal patterns and event-driven surge magnitudes for each category, from high-volume three-wheelers to lower-volume six-wheelers, demonstrating its ability to generalize across different scales and behaviors.

VII. FUTURE WORK

This research establishes a strong foundation, and future work can address current limitations and expand capabilities:

- **Real-World Data Validation:** Crucially, apply the developed methodology (feature engineering and LightGBM model) to comprehensive, integrated real-world urban traffic datasets. This will involve robustly handling noise, missing values, and integrating heterogeneous data sources for practical applicability.
- **Enhanced Synthetic Generator & Spatial-Temporal Forecasting:** Refine the generator with advanced network interactions (e.g., queue dynamics, agent-based modeling) and extend forecasting to multiple interdependent locations utilizing spatio-temporal models (e.g., Graph Neural Networks).
- **Incident Prediction Integration:** Develop and integrate probabilistic models for accident occurrence and dynamic impact profiles into the main forecasting pipeline for more robust predictions under disruptive conditions.
- **Comparative Deep Learning Analysis:** Conduct detailed comparative studies against state-of-the-art Deep Learning architectures (e.g., advanced LSTM/Transformer models) to benchmark performance on this complex problem.

VIII. REFERENCES

[1] Y. Ge, H. Liu, Y. Wang, and J. Zhang, "Short-Term Traffic Speed Forecasting Using a Deep Learning Method Based on Multitemporal Traffic Flow Volume," IEEE Access, vol. 10, pp. 82384–82395, 2022.

[2] D. Wang, H. Yu, Y. Lv, and J. Wu, "Transformer-Based Short-Term Traffic Forecasting Model Considering Spatiotemporal Dependencies," Frontiers in Neurorobotics, vol. 16, Article 917244, 2022.

[3] Y. Jia, X. Ma, L. Luo, and J. Wu, "Short-Term Traffic Flow Forecasting Considering Rainfall Impact: A Deep Learning Approach," Journal of Intelligent Manufacturing, vol. 31, pp. 1325–1337, 2020.

[4] J. C. Herrera, D. B. Work, R. Herring, X. J. Ban, Q. Jacobson, and A. M. Bayen, "Synthetic Data as a Contribution

in Multi-Agent Urban Traffic Forecasting," Transportation Research Part C, vol. 114, pp. 540–556, 2020.
[5] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, "Long Short-Term Memory Neural Network for Traffic Speed Prediction," Transportation Research Part C, vol. 54, pp. 187–197, 2018.

[6] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 3146–3154, 2017.

[7] B. M. Williams, "Forecasting Urban Travel: Past, Present and Future," Journal of the Transportation Research Board, vol. 1831, pp. 183–190, 2003.

[8] L. Vanajakshi and L. R. Rilett, "Support Vector Machine Technique for the Short-Term Prediction of Traffic Parameters," Transportation Research Record, vol. 1935, pp. 111–121, 2004.

[9] J. Zhang, Y. Zheng, and D. Qi, "Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction," AAAI Conference on Artificial Intelligence, vol. 31, no. 1, pp. 1655–1661, 2018.

[10] M. Dunne, B. Ghosh, and M. O'Mahony, "Weather Impacts on Urban Traffic," IEEE Transactions on Intelligent Transportation Systems, vol. 11, no. 2, pp. 387–391, 2010.

[11] D. Pavlyuk and O. Ivanov, "Realistic Traffic Data Generation for Machine Learning in Urban Mobility: A Simulation-Based Approach," Transportation Research Procedia, vol. 71, pp. 58–69, 2024.

[12] E. Gonzalez and D. Smith, "Using Simulation for Feature-Aware Synthetic Traffic Forecasting," Simulation Modelling Practice and Theory, vol. 136, Article 102811, 2024.

[13] K. Zhou and H. Wang, "Comparative Evaluation of Gradient Boosting Algorithms for Travel Time Forecasting," Applied Sciences, vol. 10, no. 5, pp. 1802, 2020.