Advanced Delivery Time Prediction System Using Machine Learning and Real-time Data Integration

Aishwarya S. Pathak
Assistant Professor
Dept. of Electronics and Telecommunication Engineering
BVCOEL, Pune

Pramod G. Rahate
Associate Professor
Dept. of Mechanical Engineering
BVCOEL, Pune

Neeraj A. Gangurde
Assistant Professor
Dept. of Civil Engineering
BVCOEL, Pune

Megha A. Patil
Assistant Professor
Dept. of Computer Engineering
BVCOEL, Pune

Urvashi T. Bhat
Assistant Professor
Dept. of Electronics and Telecommunication Engineering
BVCOEL, Pune

Shruti H Gunjotikar
Assistant Professor
Dept. of Electronics and Telecommunication Engineering
BVCOEL, Pune

Abstract:**Accurate delivery time prediction has become crucial in today's e-commerce and logistics environment. Traditional methods often fail to account for real-time factors, leading to customer dissatisfaction and operational inefficiencies. This paper presents an advanced delivery time prediction system that combines machine learning algorithms with real-time data integration. Using the CatBoost algorithm, known for its efficient handling of categorical data and robust performance in gradient boosting tasks, and incorporating traffic and weather data, the system achieves significantly improved prediction accuracy. CatBoost was selected over other algorithms due to its ability to handle categorical variables without extensive preprocessing, its faster training times, and its superior accuracy in complex, real-world scenarios. The experimental results show a Mean Absolute Error (MAE) of 3.2 minutes, demonstrating the system's effectiveness in real-world scenarios. The proposed system not only enhances prediction accuracy but also offers scalability for various logistics applications.**

Keywords Machine Learning, Delivery Time Prediction, CatBoost, Real-time Data Integration

I. INTRODUCTION

The exponential growth of e-commerce and food delivery services has created an urgent need for accurate delivery time predictions. Customers increasingly demand precise, real-time estimates for their orders, particularly in sectors such as food delivery, where timely service is essential. Delays or inaccurate delivery time estimates lead to frustration, reduced customer loyalty, and operational inefficiencies.

Traditional methods of estimating delivery times are often inconsistent and fail to account for dynamic factors such as traffic conditions, weather changes, and varying delivery demands. These limitations result in suboptimal resource allocation, ineffective route planning, and delayed deliveries. For example, during peak traffic hours, traditional estimation systems can underestimate delivery times by 15-20 minutes, leading to customer dissatisfaction and increased operational costs. Additionally, adverse weather conditions such as

heavy rainfall or snowstorms can cause unpredictable delays that static models fail to address effectively. The logistics industry, particularly in food and e-commerce, has struggled to address these unpredictable variables, leading to inefficiencies in driver scheduling and delivery routes.

This research aims to address these challenges by developing a robust system that leverages machine learning and real-time data integration. By combining historical delivery data with live inputs such as traffic density and weather conditions, the proposed system dynamically adjusts predictions, resulting in improved accuracy. This approach not only enhances customer satisfaction but also streamlines operational processes for logistics companies.

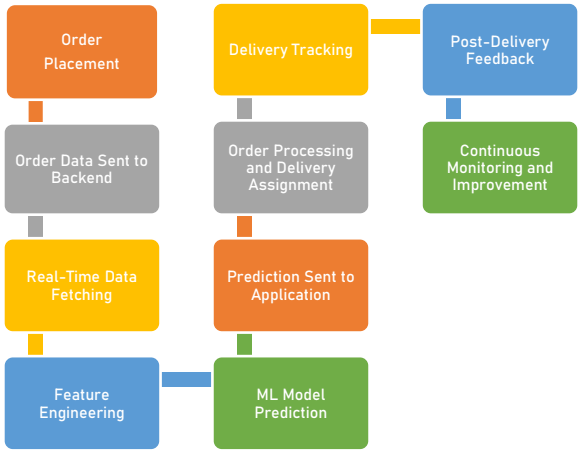


Fig 1: System Flow Diagram

II. LITERATURE SURVEY

These studies collectively underscore the critical need for integrating real-time data with machine learning models to enhance delivery time predictions. While existing research has laid a strong foundation, gaps in adapting to dynamic factors like traffic and weather remain prominent. This project addresses these shortcomings by combining real-time adaptability with advanced algorithms such as CatBoost, offering a more robust and scalable solution for the logistics industry.

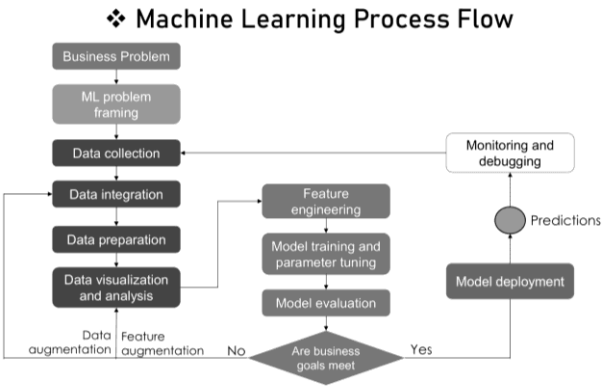
Several researchers have explored the potential of machine learning in logistics and delivery time prediction. Below are notable contributions:

1. **Prediction of service time for home delivery services using machine learning** (Jan Wolter et al., 2024): This study utilized artificial neural networks (ANN) to predict delivery service times but did not address real-time dynamic factors such as traffic and weather.
2. **Machine Learning-Based Traffic Flow Prediction and Intelligent Traffic Management** (Zheng Xu et al., 2024): The authors proposed a multiview spatiotemporal convolutional model for traffic flow prediction, emphasizing the importance of real-time traffic data.
3. **Predictive Analytics for Real-Time Supply Chain Agility** (Abeer Aljohani, 2023): This work combined predictive analytics with machine learning, highlighting the benefits of AI in supply chain risk management. However, it lacked a focus on delivery time prediction.
4. **Data-Driven Optimization for Last-Mile Delivery** (J. Pan et al., 2021): This paper introduced machine learning techniques combined with vehicle routing optimization but did not integrate real-time traffic or weather inputs.

These studies highlight gaps in real-time adaptability and integration, which this research aims to address.

III. CONCEPTUAL REVIEW OF METHODS

The conceptual framework for this research focuses on three main components: the selection of machine learning models, integration of real-time data, and advanced feature engineering. These elements collectively form the foundation of the proposed delivery time prediction system.



A. Machine Learning Models

Fig 2: Machine Learning Process Flow

The system employs a range of machine learning models, each contributing unique strengths to optimize prediction accuracy:

1. **Linear Regression:** This model establishes baseline relationships between input features and delivery times, offering a clear benchmark. Its simplicity makes it computationally efficient, but it struggles with capturing non-linear interactions, limiting its real-world applicability. Despite this, Linear Regression serves as a valuable reference point for evaluating the performance of more advanced models.
2. **Random Forest:** An ensemble learning method that combines multiple decision trees to improve prediction accuracy. Random Forest excels in handling non-linear relationships and is robust to overfitting. It also provides insights into feature importance, aiding in understanding the variables influencing delivery times.
3. **CatBoost:** This advanced gradient boosting algorithm is specifically designed to handle categorical data efficiently, eliminating the need for extensive preprocessing. It offers faster training times and achieves superior accuracy compared to traditional boosting algorithms. CatBoost's ability to adapt to complex, real-world scenarios makes it the best-performing model for this system.

B. Real-Time Data Integration

Real-time data forms a critical component of the system, enabling dynamic adjustments to predictions based on current external conditions. The integration involves two primary sources:

1. **Traffic Data:** Obtained through periodic API calls to Google Maps API, this data includes live traffic density, congestion patterns, and estimated travel times. These metrics are dynamically incorporated into the model, allowing for accurate adjustments in high-traffic scenarios. For example, during peak hours, the system accounts for delays caused by congestion, enhancing prediction reliability.
2. **Weather Data:** Sourced from OpenWeatherMap API, this data includes real-time weather conditions such as precipitation, wind speed, and temperature. The system assigns delay weights to adverse conditions like storms, ensuring that weather

impacts on delivery times are accurately reflected. By leveraging asynchronous API calls, the backend architecture ensures minimal latency in fetching and preprocessing this data.

C. Feature Engineering

Effective feature engineering is central to the success of the system, transforming raw data into meaningful inputs for the machine learning models:

1. **Distance Calculation:** The Haversine formula computes the shortest distance between the restaurant and the delivery location based on latitude and longitude coordinates. This ensures precise distance measurements, a key factor in estimating delivery times.
2. **Time-based Features:** Temporal patterns, such as the hour of the day, day of the week, and holiday indicators, are extracted to capture trends in delivery times. For instance, deliveries during weekends or peak hours are typically delayed, and these insights are encoded into the model.
3. **Weather and Traffic Impacts:** Features reflecting real-time conditions, such as traffic density scores and quantified weather impacts, are engineered to enhance the model's adaptability. For example, high traffic density during rush hours or adverse weather conditions like heavy rain are assigned higher delay weights to ensure accurate predictions.

Together, these components create a robust and scalable system capable of delivering highly accurate predictions while adapting dynamically to real-world conditions.

IV. SYSTEM IMPLEMENTATION

Fig 3: System Architecture: ER diagram

A. Data Processing Pipeline

1. **Historical Data Preprocessing:** Handling missing values and outliers, followed by normalization of numeric features.
2. **Real-Time Data Integration:** Continuous fetching of live data from traffic and weather APIs.
3. **Feature Extraction:** Generating derived features like traffic density scores and weather impact metrics.

B. Model Training and Evaluation

1. **Cross-Validation:** Ensuring model generalization through K-fold cross-validation.
2. **Hyperparameter Tuning:** Optimizing parameters like learning rate, depth, and tree iterations for CatBoost.
3. **Performance Metrics:** Evaluating models based on MAE, RMSE, and R-squared values.

V. RESULTS AND ANALYSIS

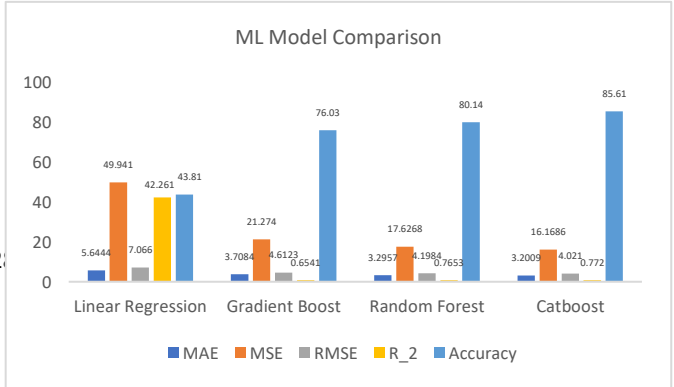


Fig 4. Machine Learning Model Comparison.

Model	MAE (minutes)	RMSE (minutes)	R-squared
Linear Regression	5.64	7.07	0.42
Random Forest	3.29	4.19	0.76
Gradient Boosting	3.70	4.60	0.65
CatBoost	3.20	4.00	0.77

A. Prediction Accuracy

The system achieved an MAE of 3.2 minutes, RMSE of 4.0 minutes, and R-squared of 0.77, outperforming traditional methods. These metrics underline the model's ability to handle real-world complexities. The table above highlights the comparative performance of various models, showcasing CatBoost's superiority in predictive accuracy.

B. Real-World Performance

Case studies demonstrate the system's adaptability:

1. **Urban Delivery During Peak Traffic:** Accurate predictions within 8 minutes of actual times.
2. **Adverse Weather Conditions:** Effective estimation with a 10-minute error margin during storms.
3. **Multiple Deliveries:** Accurate adjustments for scenarios involving multiple simultaneous orders.

VI. CONCLUSIONS

The proposed system demonstrates significant advancements in delivery time prediction by integrating machine learning with real-time data. The CatBoost algorithm, coupled with comprehensive feature engineering, achieves superior accuracy and scalability. This research has the potential to revolutionize the logistics industry by enabling more efficient resource allocation, reducing operational costs, and enhancing customer satisfaction. By addressing critical real-world challenges like unpredictable traffic and adverse weather conditions, this system can significantly improve delivery reliability. Future work will explore enhanced route optimization, additional data sources, and expansion to broader logistics domains, setting the foundation for a smarter and more adaptive logistics ecosystem.

VII. REFERENCES

Wolter, J., & Hanne, T. (2024). Prediction of service time for home delivery services using machine learning. Soft Computing, 28, 5045–5056.

Xu, Z., Yuan, J., Yu, L., Wang, G., & Zhu, M. (2024). Machine learning-based traffic flow prediction and intelligent traffic management. International Journal of Computer Science and Information Technology, 2(1).

Aljohani, A. (2023). Predictive analytics and machine learning for real-time supply chain risk mitigation and agility. Sustainability, 15(20).

Pan, J., Li, M., & Chen, W. (2021). Data-driven optimization for last-mile delivery in online food delivery platforms. Journal of Transportation Technology.

Liu, Z., et al. (2021). Data-driven optimization for last-mile delivery. Complex & Intelligent Systems.

Thomas, A., & Panicker, V. V. (2023). Application of machine learning algorithms for order delivery delay prediction in supply chain disruption management. SpringerLink.

Bhalla, S., et al. (2023). Case study on delivery time determination using a machine learning approach in small batch production companies. Journal of Intelligent Manufacturing.

Google Maps API: Traffic and directions API. Available at: https://developers.google.com/maps/documentation.

OpenWeather Map API: Weather data API . Available at: https://openweathermap.org/api.