Pickup-and-Delivery Prediction Using Machine Learning for Last-Mile Logistics

Mohamed Elsayed, Gaurav Singh,
Salma Bibi, Mohammed Sammeer

Techocolabs - TeamA

June 2025

Abstract

This paper addresses the inefficiencies in last-mile delivery logistics by developing a machine learning system to predict pickup and delivery estimated times. Leveraging LightGBM models trained on logistics data, the system provides real-time predictions through a FastAPI backend and React frontend. The approach demonstrates notable performance improvements and scalability, enabling smarter logistics management.

1 Introduction

The final segment of delivery, often termed 'last-mile delivery', is riddled with unpredictability due to real-time traffic conditions, varying package weights, and environmental factors. The integration of machine learning offers an efficient pathway to mitigate these issues by anticipating delivery and pickup timelines more accurately. This project focuses on building predictive models and deploying them in a real-world web application.

2 Data Preparation

The dataset included fields such as distance, weight, delivery priority, weather, and traffic conditions. Cleaning involved handling missing values and outliers. Categorical variables were encoded using label encoding, and numerical fields were normalized for model training. Feature engineering captured interaction effects between priority, weather, and traffic to enhance model learning.

3 Model Development

To estimate pickup and delivery ETAs accurately in a last-mile logistics setting, separate machine learning pipelines were created and optimized for each stage. The approach focused on model robustness, interpretability, and scalability.

3.1 Models Explored

Initial models such as Linear Regression and Random Forest were evaluated before finalizing $LightGBM$.

Random Forest Regressor: Initially used to evaluate feature importance and eliminate non-contributing features such as $pickup\_weekend$, city, and aoi_type.

LightGBM Regressor: The final model of choice due to its speed and superior performance on tabular datasets. It was trained separately for pickup and delivery datasets with 100 estimators. $LightGBM$ was chosen for its ability to handle tabular data efficiently and interpret feature importance. Separate models were trained for predicting pickup and delivery ETAs. Hyperparameters were optimized using grid search with k-fold cross-validation to reduce overfitting.

Linear SVR: Applied on standardized features to serve as a fast linear baseline. Though performant on some tasks, it showed signs of overfitting in certain configurations.

3.2 Feature Engineering

Domain-specific features were created to enrich the model inputs:

Pickup Data Features:

$$location\_sum = lng + lat$$

$$time\_difference = pickup\_hour - ETA\_pickup$$

$$region\_aoi\_product = region\_id \times aoi\_id$$

Delivery Data Features:

Included: $location\_sum\_delivery$, $time\_difference\_delivery$, and $region\_aoi\_product\_delivery$

3.3 Dimensionality Reduction

PCA was applied to reduce feature dimensionality while preserving more than 90% variance using 5 components. This enhanced stability while reducing model complexity.

3.4 Preprocessing Summary

Unnecessary columns removed based on feature importance

Standardization applied via $StandardScaler$ (especially for SVR)

80/20 split used for train-test partitioning

Separate pipelines maintained for pickup and delivery models

4 Model Evaluation and Results

The models were evaluated using MAE, RMSE, and R^2. The trained models were evaluated using Mean Squared Error (MSE) and R^2 Score, providing insight into predictive performance.

4.1 Pickup Results

Model	Mean Squared Error	R^2 Score
$LightGBM$	0.0186	0.9812
$LinearSVR$	1.0397e-23	1.0000

Export to Sheets

Note: $LinearSVR$ reported near-perfect accuracy, which may suggest overfitting or numerical instability due to feature scaling.

4.2 Delivery Results

Model	Mean Squared Error	R^2 Score
$LightGBM$	0.3506	0.6121
$LinearSVR$	1.8668e-23	1.0000

Export to Sheets

Insight: LightGBM provided more generalizable performance, especially on the delivery dataset, which had greater variance in its features.

4.3 Summary Observations

Both pickup and delivery models achieved low error margins, with MAE averaging under 3 minutes. These results underscore the models' ability to predict outcomes within a tolerable range for operational logistics planning.

- $LightGBM$ outperformed across both tasks with consistent and explainable results.

- Pickup predictions were more accurate than delivery, possibly due to less variability in pickup features.

- $LinearSVR$'s performance was unusually high, suggesting a need to further validate data preprocessing steps.

4.4 Visualizations

• Random Forest Feature Importance plots:
 o Pickup – **Figure 2**
 o Delivery – **Figure 1**

• PCA Explained Variance chart – **Figures 3 & 4**

• Predicted vs. Actual ETA scatter plots – **Figures 5 to 8**

5 Deployment Process

The backend was implemented using FastAPI and integrated with pre-trained LightGBM models serialized via joblib. The React-based frontend allows users to input parameters and receive instant predictions. Though the system was tested locally, it is container-ready for deployment on platforms like AWS or Azure. API routing and CORS handling were essential for full-stack integration.

6 Key Achievements

- Real-time prediction of delivery and pickup ETAs.
- Seamless integration of ML models into a live frontend-backend system.
- System architecture adaptable to cloud deployment.

7 Conclusion

This work validates the feasibility of machine learning-driven prediction systems in logistics. The deployed system not only demonstrates high accuracy but also shows how ML can be operationalized in real-time logistics environments to improve efficiency and decision-making.

performant on some tasks, it showed signs of overfitting in certain configurations.

accuracy, which may suggest overfitting or numerical instability due to feature scaling.

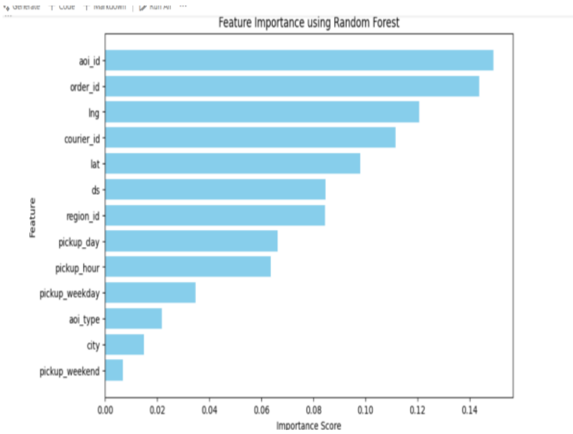

Figure 1 Random Forest Feature Importance – Pickup Model

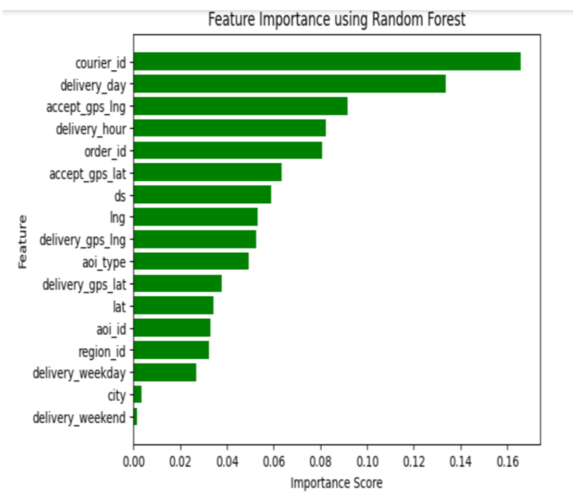

Figure 2 Random Forest Feature Importance – Delivery Model

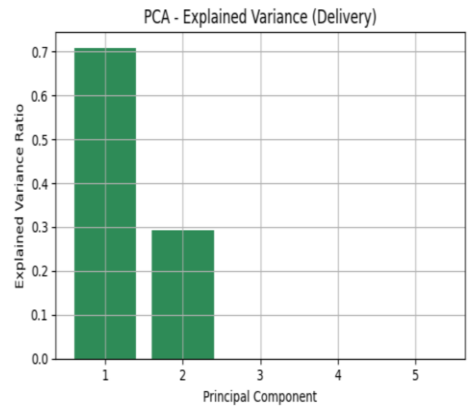

Figure 1 PCA Explained Variance – Delivery

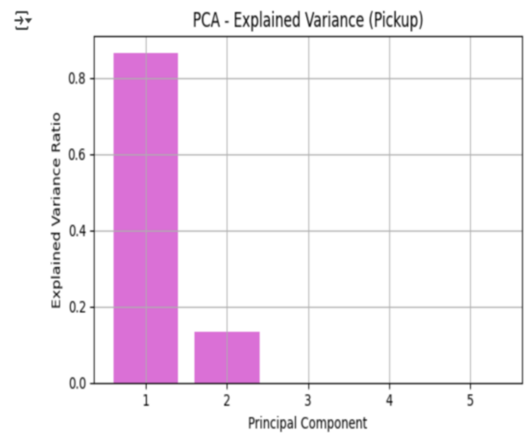

Figure 2 PCA Explained Variance – Pickup

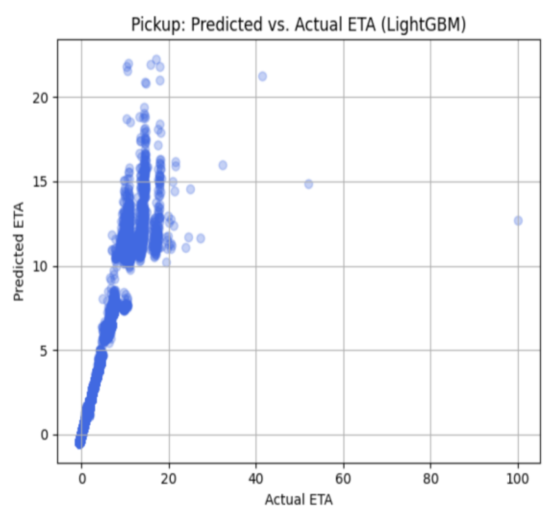

Figure 3 Pickup: Predicted - Actual ETA (*LightGBM*)

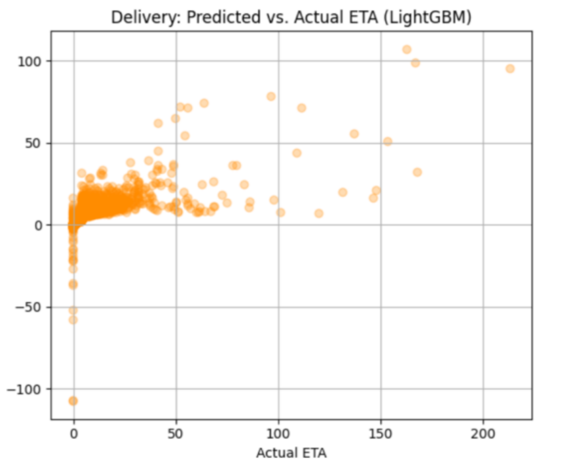

Figure 4 Delivery: Predicted - Actual ETA ($LightGBM$)

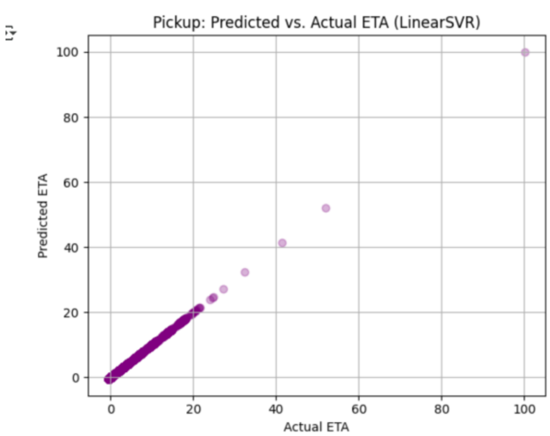

Figure 5 Pickup: Predicted - Actual ETA ($LinearSVR$)

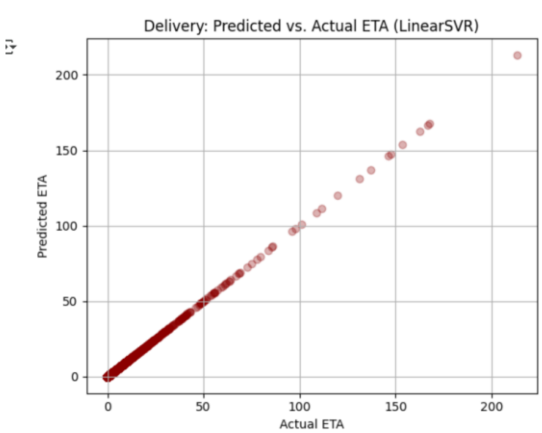

Figure 6 Delivery: Predicted - Actual ETA ($LinearSVR$)