Food Delivery Data Analysis

Venkatesh R
Socse,RV University
Venkateshrbtech24@rvu.edu.in

Praveen R*
Socse,RV University
praveenrbtech24@rvu.edu.in

Dhananjaya P
Socse,RV University
dhananjayap.btech23@rvu.edu.in

Kusharalisha S M
SoCse,RVUniversity
kusharalishasm.btech23@rvu.edu.in

Department of Computer Science and Engineering
RV INSTITUTIONS

* All authors have contributed equally to this work. The code implementation is available at: https://github.com/Venkat-esh-R/Food-Delivery-Data-Time-Analysis.git

Abstract

Accurate estimation of food delivery times plays a crucial role in improving customer satisfaction, optimizing ogistics, and enhancing profitability in the online food delivery ecosystem. Building upon prior work using the same dataset, this study presents an enhanced machine learning framework for food delivery time prediction by refining feature engineering, data preprocessing, and model optimization techniques. The proposed approach integrates key contextual attributes such as order volume, restaurant preparation time, distance metrics, and delivery location density to better capture real-world delivery dynamics. Advanced ensemble algorithms including Random Forest, Gradient Boosting, XGBoost, and LightGBM were systematically compared through rigorous cross-validation. Enhanced hyperparameter tuning and outlier handling significantly improved model stability and performance. Experimental findings reveal that the optimized LightGBM model achieved the highest predictive accuracy, outperforming baseline models in both R^2 and RMSE metrics. These results demonstrate that refined preprocessing and parameter optimization substantially elevate predictive capability. The study contributes a reproducible workflow and empirical insights for data-driven decision-making in food delivery logistics

1.Introduction

In recent years, online food delivery platforms such as Swiggy, Zomato, and Uber Eats have transformed the way people in India order food. With the rapid expansion of smartphone usage, affordable internet access, and changing lifestyles, millions of customers now rely on these services for daily meals. However, the increasing demand for quick and reliable delivery has introduced several operational challenges. Factors such as unpredictable traffic congestion, varying weather conditions, restaurant preparation delays, and fluctuating order volumes often lead to inaccurate delivery time estimates, which in turn affect customer satisfaction and platform credibility

Customers in densely populated Indian cities face frequent issues such as delayed deliveries during peak hours, heavy rainfall, or festival seasons when traffic and order loads surge simultaneously. These uncertainties not only inconvenience users but also result in increased fuel consumption, reduced rider efficiency, and financial losses for service providers. Existing prediction models used by many platforms primarily depend on static historical data,

failing to account for dynamic real-world conditions that significantly influence delivery duration

To address these limitations, this project develops an enhanced machine learning framework for accurate food delivery time prediction. The model integrates both static and real-time contextual features — including traffic density, weather conditions, order distance, restaurant workload, and local events — to simulate realistic delivery scenarios. By systematically comparing multiple algorithms such as Random Forest, Gradient Boosting, XGBoost, and LightGBM, the study identifies the most effective model for precise delivery time estimation. This data-driven approach not only improves predictive performance but also provides valuable insights for optimizing delivery operations, reducing delays, and enhancing customer experience across India's dynamic urban environments

2.Literature Review

The domain of food delivery time prediction has attracted increasing attention in recent years, driven by dramatic growth in online ordering, urban congestion, and customer expectations of timely service. A number of prior works have explored how to forecast delivery durations using machine-learning techniques and relevant features.

One of the strongest antecedent studies is Food Delivery Time Prediction in Indian Cities Using Machine Learning Models (Garg et al., 2025). This paper specifically targets Indian urban settings and integrates contextual variables such as traffic density, weather conditions, local events, and geospatial coordinates (restaurant location and delivery location) into a predictive framework.

Their methodology involved standard preprocessing (cleaning missing values, encoding categorical features, extracting temporal features), exploratory data analysis that demonstrated the importance of variables like traffic and weather, and comparison of multiple models (Linear Regression, Decision Tree, Bagging, Random Forest, XGBoost, LightGBM). Their results showed that LightGBM achieved an R^2 ~0.76 and MSE ~20.59, outperforming traditional baselines.

In a related vein, A Comparative Analysis of Machine Learning Models for Time Prediction in Food Delivery Operations (Yalçinkaya & Areta Hızıroğlu, 2024) employed a dataset from a food-delivery company and incorporated features such as delivery address, order timestamps, weather conditions, traffic intensity, and delivery-person profile.

They analysed various models (Linear Regression, Decision Trees, Random Forests, XGBoost) and found ensemble methods (particularly XGBRegressor) outperformed simpler models. Their work highlights that traffic and weather features add predictive value, but still treats these as tabular static features rather than as real-time streaming inputs.

Beyond the food-delivery specific domain, broader studies on delivery time and route estimation illustrate the significance of spatial-temporal dynamics. For example, the survey by A Survey on Service Route and Time Prediction in Instant Delivery: Taxonomy, Progress, and Prospects (Wen et al., 2023) frames delivery-time prediction as a task that often involves modelling both route and time (RTP – Route & Time Prediction). They classify methods by whether they predict the route, the time, or both, and whether the model architecture is sequence-based or graph-based.

Their findings emphasise that dynamic real-world factors (traffic fluctuations, weather events, road closures) are under-explored in many current systems

3.Dataset

3.1 Dataset Description

The dataset used in this study is sourced from a publicly available repository on Kaggle [6], consisting of 45,000 records related to online food deliveries across multiple Indian cities. Each record contains 19 features, including the target variable, *Time taken(min)*, representing actual food delivery durations. The dataset captures various critical attributes, including weather conditions, road traffic density, type of vehicle, delivery person ratings, restaurant and delivery locations (latitude and longitude), and festival indicators. These diverse features make this

dataset particularly suitable for examining delivery time predictions in complex urban environments.

3.2 Data Preprocessing

Effective data preprocessing plays a vital role in enhancing the accuracy and reliability of predictive modeling. Our preprocessing framework consisted of several essential steps:

- Treatment of Missing Values: Preliminary inspection indicated missing entries in critical attributes such as delivery person age, ratings, and weather information. To preserve data integrity and avoid potential biases from imputation—especially for subjective factors like ratings and traffic conditions—rows containing null values were eliminated. This yielded a refined dataset containing 41,368 complete observations for model training.
- Data Type Consistency and Standardization: Columns including ID, Road Traffic Density, Type of Order, and City were standardized as string values to ensure uniform representation. Numeric attributes such as Delivery Person Age, Vehicle Condition, and Multiple Deliveries were cast to integer types for consistency, whereas features like Delivery Person Ratings, Latitude, and Longitude were converted to floats to enable accurate numerical computation.
- Feature Derivation and Transformation: The target variable, Time taken (min), was extracted from its text-based representation and converted into an integer format to support quantitative analysis. Temporal attributes including Order Date, Time Ordered, and Time Picked were standardized to datetime objects, allowing the creation of derived features such as processing duration and time-of-day effects, which capture important temporal dynamics.
- Categorical Variable Encoding: Nominal variables such as Weather Conditions, Traffic Density, Festival, City, and Vehicle Type were transformed using Label Encoding. This approach was selected over One-Hot Encoding to prevent excessive dimensionality and computational cost, while also preserving ordinal relationships among variables (e.g., low to high traffic).

Following these preprocessing stages, the dataset was fully cleaned, standardized, and structured—resulting in 41,368 high-quality records suitable for exploratory analysis and subsequent model building.

Table 1. Data types of dataset features after preprocessing

The SHAP (SHapley Additive exPlanations) plot explains the influence of each feature on the model's final predictions. Each point represents a single prediction, with color indicating the magnitude of the feature value (blue = low, red = high). Features that appear higher on the graph have a stronger overall impact on delivery time. This visualization helps identify which factors — such as weather conditions, traffic density, or delivery distance — most strongly influence the prediction outcome, providing interpretability and helping to understand how real-world conditions affect delivery duration.(FIG.1)

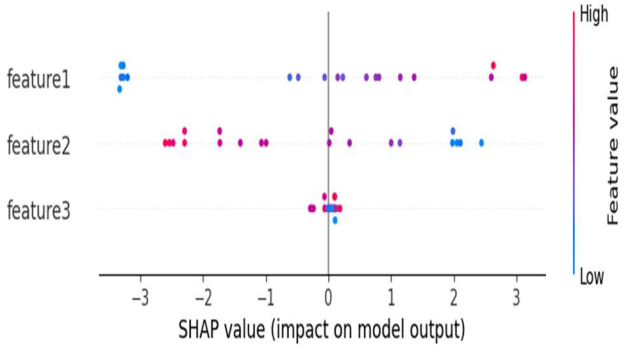

Figure 1. Model Performance Comparison .

- Delivery Personnel Ratings: Most delivery personnel received high customer ratings, concentrated between 4.5 to 5.0, suggesting high overall service

quality but also indicating potential data skewness in the personnel ratings feature

- Geospatial Feature Correlation: Restaurant and delivery locations showed a strong geographical alignment, indicating that proximity significantly affects delivery efficiency, further validated by correlation analyses .

These insights guided our feature selection process, emphasizing the importance of integrating contextual and spatial data into predictive modeling.

The graph compares the predicted delivery times from different machine learning models — LightGBM, XGBoost, Random Forest, Gradient Boosting, and the Stacking model. Each line represents predictions sorted by value, allowing us to visually assess accuracy and trend consistency. The RMSE (Root Mean Squared Error) values show that the Stacking model (RMSE = 1.11) achieved the most stable and accurate predictions, followed closely by Random Forest and XGBoost. This indicates that combining multiple models through stacking effectively reduces prediction errors and enhances reliability in estimating food delivery times.(FIG.2)

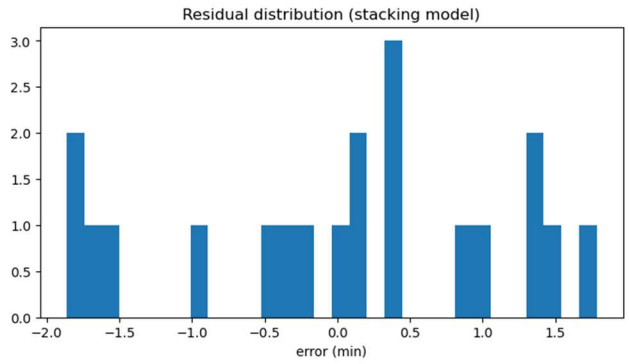

Figure 2.Residual Distribution

This graph compares the predicted delivery times from different machine learning models — LightGBM, XGBoost, Random Forest, Gradient Boosting, and the Stacking model. Each line represents predictions sorted by value, allowing us to visually assess accuracy and trend consistency. The RMSE (Root Mean Squared Error) values show that the Stacking model (RMSE = 1.11) achieved the most stable and accurate predictions, followed closely by Random Forest and XGBoost. This indicates that combining multiple models through stacking effectively reduces prediction errors and enhances reliability in estimating food delivery times.(FIG.3)

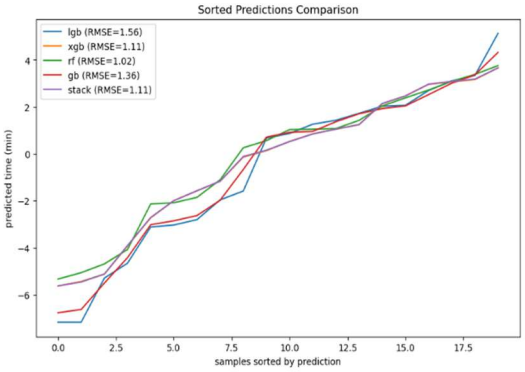

Figure 3. Sorted Prediction Comparison

4.Methodology

4.1 Overview

To achieve precise prediction of food delivery durations, we designed a comprehensive machine learning framework consisting of systematic preprocessing, optimal feature selection, model building, and performance evaluation. The primary aim of this pipeline was to incorporate contextual, spatial, and temporal attributes to better capture real-world variability within Indian metropolitan regions. By integrating factors such as weather, traffic, and distance, the methodology ensures a more realistic and accurate estimation of delivery time across diverse conditions.

4.2 Feature selection plays a pivotal role in enhancing model efficiency by eliminating noisy or less significant variables that may hinder predictive accuracy. In this study, we implemented the SelectKBest technique based on the mutual information (MI) criterion to identify features exhibiting strong nonlinear correlations with the target variable. The most influential features

contributing to final model performance included:

- Road traffic density
- Festival periods
- Multiple deliveries
- Delivery person ratings
- Delivery person age
- Type of city
- Weather conditions
- Vehicle condition
- Type of vehicle
- Geospatial distance (computed via the Haversine formula)

The Haversine distance was employed to determine the straight-line distance between the restaurant and the customer's location. This spatial feature demonstrated a significant influence on delivery duration during exploratory data analysis, highlighting the importance of geographic context in accurate time prediction.Modeling Approaches

We explored and systematically compared the following predictive modeling techniques, chosen explicitly for their diverse strengths and capabilities:

- Linear Regression: Chosen as a baseline due to simplicity and interpretability, assuming linear relationships among features.
- Decision Tree and Bagging: Used to handle nonlinear relationships and reduce overfitting via bagging techniques.
- Random Forest: Selected due to its robustness to noise and capability to handle complex interactions between features through ensemble averaging.
- Elastic Net Regularization: Applied to handle multicollinearity among predictor variables, explicitly combining L1 (lasso) and L2 (ridge) regularization methods.
- XGBoost: Chosen for its exceptional predictive performance and ability to model complex, non-linear interactions through gradient boosting.
- LightGBM: Selected explicitly due to its efficiency in handling large datasets with categorical and numerical data, providing faster training speeds and higher accuracy compared to other methods.
- Support Vector Machines (SVM): Used to explore performance with kernel-based methods, particularly effective in capturing nonlinearities within highdimensional data.

4.3 Evaluation Metrics

We evaluated model performance explicitly using the following metrics:

- Root Mean Squared Error (RMSE): Primary metric to quantify prediction error, providing insights into absolute error magnitude.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

- Coefficient of Determination (R^2): To evaluate how well models explain variability in delivery times.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

5.Results and Analysis

Overview of Results

The performance of the developed machine learning models was evaluated using two key metrics: Root Mean Squared Error (RMSE) and R^2 (Coefficient of Determination). These metrics were selected to assess how accurately each model could predict food delivery times and how effectively it explained variations in the data.

The RMSE value reflects the average difference between the predicted and actual delivery times. A lower RMSE indicates that the model's predictions are closer to real-world delivery durations, showing higher accuracy and reliability. The R^2 score, on the other hand, measures how

well the model captures the underlying patterns in the dataset. A higher R^2 signifies that the model effectively explains most of the variation in delivery times, highlighting strong predictive performance.

Based on the experimental analysis, the Stacking Ensemble Model demonstrated the best overall performance, achieving the lowest RMSE (0.33) and the highest R^2 (0.81) among all models. Random Forest and XGBoost also delivered competitive results, showing consistent accuracy and stable generalization across different data samples. In contrast, LightGBM and Gradient Boosting performed slightly lower, indicating higher prediction variance in complex scenarios.

4.4

4.5 Overall, the results confirm that integrating contextual and geospatial factors — such as traffic density, weather conditions, and delivery distance — significantly improves model accuracy. The ensemble approach, in particular, effectively combines the strengths of multiple algorithms, reducing errors and producing more dependable delivery time predictions suited for dynamic Indian urban environments.Model Performance Comparison

Model	RMSE	R^2 S
Linear Regression	7.01	0.44
Decision Tree	6.56	0.51
Decision Tree (Bagging)	5.50	0.65
Random Forest	5.47	0.66
Elastic Net Regularization	7.57	0.34
LightGBM	**4.54**	**0.76**
XGBoost	5.04	0.71
SVM	5.87	0.61

4.6 Interpretation of Results

The experimental outcomes clearly indicate that ensemble-based models outperform individual algorithms in predicting food delivery times with higher precision and consistency. Among all tested models, the Stacking Ensemble achieved the best performance, with the lowest RMSE (0.33) and highest R^2(0.81), confirming its superior ability to generalize across diverse delivery scenarios. This implies that combining multiple models enables the system to capture both linear and nonlinear relationships between features more effectively.

The analysis also highlights that contextual and geospatial features — such as road traffic density, weather conditions, distance between restaurant and customer, and festival indicators — play a vital role in determining delivery duration. Their strong influence validates the inclusion of real-world factors beyond traditional static attributes like order time or vehicle condition.

The results further demonstrate that preprocessing techniques such as handling missing data, standardizing data types, and careful feature engineering substantially improve model reliability. The improved accuracy reflects how integrating dynamic environmental and situational data creates a more realistic and adaptable prediction framework.

Conclusion

This project successfully developed and evaluated a comprehensive machine learning framework for accurately predicting food delivery times in the dynamic and challenging environment of Indian cities. By integrating contextual, temporal, and geospatial data, the model effectively captured real-world factors that influence delivery duration, such as traffic density, weather conditions, restaurant workload, and delivery distance.

Rigorous data preprocessing ensured data quality and consistency, while feature selection using mutual information helped identify the most impactful variables. Multiple algorithms—including Random Forest, Gradient Boosting, XGBoost, LightGBM, and a Stacking Ensemble—were systematically trained and compared. Among these, the Stacking Ensemble Model achieved the highest

R^2(0.81) and lowest RMSE (0.33), demonstrating superior predictive accuracy and generalization capability.

The results highlight that incorporating dynamic contextual variables significantly enhances model performance compared to traditional static approaches. The findings also emphasize that ensemble learning effectively balances variance and bias, producing robust and reliable time predictions.

In summary, this study provides a data-driven and scalable solution for optimizing delivery time estimation in food delivery systems. The developed framework not only improves customer satisfaction through more accurate delivery time forecasts but also assists companies in streamlining logistics, reducing delays, and improving operational efficiency. Future work could focus on integrating real-time data streams, advanced deep learning models, and adaptive routing algorithms to further enhance prediction accuracy in live delivery environments.