Food Delivery Time Prediction in Indian Cities Using Machine Learning Models

Ananya Garg*
ananya22068@iiitd.ac.in

Mohmmad Ayaan*
ayaan22302@iiitd.ac.in

Swara Parekh*
swara2022524@iiitd.ac.in

Vikranth Udandarao*
vikranth22570@iiitd.ac.in

Department of Computer Science
Indraprastha Institute of Information Technology, Delhi

*All authors have contributed equally to this work. The code implementation is available at: https://github.com/Vikranth3140/Food-Delivery-Time-Prediction.

Abstract

Accurate prediction of food delivery times significantly impacts customer satisfaction, operational efficiency, and profitability in food delivery services. However, existing studies primarily utilize static historical data and often overlook dynamic, real-time contextual factors crucial for precise prediction, particularly in densely populated Indian cities. This research addresses these gaps by integrating real-time contextual variables such as traffic density, weather conditions, local events, and geospatial data (restaurant and delivery location coordinates) into predictive models. We systematically compare various machine learning algorithms, including Linear Regression, Decision Trees, Bagging, Random Forest, XGBoost, and LightGBM, on a comprehensive food delivery dataset specific to Indian urban contexts. Rigorous data preprocessing and feature selection significantly enhanced model performance. Experimental results demonstrate that the LightGBM model achieves superior predictive accuracy, with an R^2 score of 0.76 and Mean Squared Error (MSE) of 20.59, outperforming traditional baseline approaches. Our study thus provides actionable insights for improving logistics strategies in complex urban environments. The complete methodology and code are publicly available for reproducibility and further research.

1. Introduction

The rapid growth of online food delivery services has significantly transformed urban consumption patterns, particularly in Indian cities where platforms like Zomato and Swiggy dominate the market. Providing accurate and reliable estimates of delivery times is essential not only for enhancing customer satisfaction but also for optimizing operational efficiency and reducing overall delivery costs. However, accurately predicting food delivery times remains challenging due to various uncontrollable and dynamic factors such as traffic congestion, variable weather conditions, and sudden demand fluctuations caused by local festivals or events.

Existing research in the domain predominantly relies on static historical data, such as historical average delivery durations and past order volumes. These traditional methods often neglect dynamic, context-specific factors like real-time traffic conditions, weather variability, and geographic complexities, which are particularly relevant in the context of Indian urban environments. The oversight of these critical variables leads to inaccurate predictions and subsequently undermines operational performance.

In this paper, we explicitly address this research gap by proposing and evaluating a novel machine learning-based predictive framework that leverages real-time contextual and geospatial information. Our approach integrates critical features such as real-time traffic density, weather conditions, and geographic distance between restaurants and customer locations, combined with comprehensive demographic and logistical information about delivery personnel and order specifics.

To achieve our objectives, we systematically evaluate and compare a range of predictive modeling techniques, including traditional methods like Linear Regression and advanced ensemble models such as Random Forest, XGBoost, and LightGBM. Through rigorous preprocessing and careful feature selection, we demonstrate that the integration of real-time contextual data significantly enhances predictive accuracy. Our empirical analysis clearly indicates the superior performance of ensemble models, particularly LightGBM, in accurately modeling the complex relationships inherent in food delivery logistics.

The primary contributions of this research include:

- Identification of crucial contextual and geospatial variables significantly impacting food delivery times in Indian cities.
- Systematic integration of dynamic real-time data, including traffic conditions and weather patterns, into predictive models.
- Comprehensive evaluation and comparison of various machine learning techniques, identifying LightGBM as the optimal approach with an R² score of 0.76.
- A publicly available reproducible implementation, fostering further research and practical applications.

The remainder of the paper is structured as follows: Section 2 presents a detailed literature review and identifies existing gaps. Section 3 provides the dataset description, data preprocessing steps, and exploratory data analysis. Section 4 discusses the detailed methodology, including model training and validation strategies. Section 5 presents and analyzes the results, followed by Section 6, which discusses practical implications, limitations, and future research directions. Finally, Section 7 summarizes our findings and contributions succinctly.

2. Literature Review

The prediction of delivery times has been extensively studied across various domains, including general logistics, e-commerce, ride-sharing, and online food delivery. Predicting delivery times accurately helps businesses enhance customer satisfaction, optimize resources, and reduce operational costs. Traditional approaches have utilized regression methods, including Linear Regression and Decision Trees, providing initial insights into relationships between predictors and delivery durations [1]. However, the complexity and variability inherent in urban delivery logistics have often rendered these basic models insufficient.

Recent studies have increasingly explored ensemble learning methods due to their improved predictive capabilities. For example, Yalçinkaya and Hiziroğlu [2] conducted a comparative analysis using machine learning models such as Random Forests and Gradient Boosting, highlighting that ensemble models consistently outperform simpler models such as Linear Regression, especially when dealing with heterogeneous datasets and complex relationships among features. Similarly, Şahin and Içen [3] demonstrated that Random Forest algorithms effectively handle the complexity of online food delivery prediction tasks by integrating real-world features like traffic density and order characteristics, achieving high accuracy rates (approximately 95%) with Random Forests. However, their approach struggled with class imbalances and did not incorporate dynamic or real-time contextual data, which is crucial for practical deployment.

Moreover, recent studies have begun exploring advanced predictive approaches in related logistics domains. For example, Chen and Guestrin introduced XGBoost [4], a scalable ensemble method widely applied in various prediction problems due to its robustness and high predictive accuracy. Similarly, Ke et al. developed LightGBM [5], an efficient gradient boosting framework optimized for handling large-scale and complex datasets with heterogeneous features. Although these advanced models offer promising accuracy improvements, their applicability specifically in the Indian urban delivery context remains under-explored, particularly regarding the integration of real-time contextual data, including weather, traffic, and local events.

A significant gap in current literature is the limited consideration given to real-time contextual and geographical features, especially within the Indian food delivery ecosystem. Indian cities are characterized by unique logistical challenges, including unpredictable traffic congestion, weather variability, high-density urban planning, frequent local events and festivals, and diverse city types (metro, urban, semi-urban), making static historical prediction methods inadequate for reliable and robust predictions.

To explicitly address these gaps, our study integrates real-time contextual factors (traffic conditions, weather data, city-specific information) and precise geospatial data into predictive models, specifically examining their impact on food delivery time predictions. We comprehensively evaluate various regression and ensemble machine learning algorithms—including Random Forest, XGBoost, and LightGBM—to identify the optimal predictive methodology suited for Indian urban conditions. Our approach is designed to enhance predictive accuracy, operational relevance, and practical usability in real-world settings, addressing crucial gaps identified in existing literature.

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 preprocessing is crucial for accurate predictive modeling. Our preprocessing pipeline involved several key steps:

- **Handling Missing Values**: Initial analysis revealed missing values in crucial features like delivery personnel age, ratings, and weather conditions. We opted to remove rows containing null values, resulting in a final cleaned dataset of 41,368 records. This choice was made to ensure reliability of model training, as imputation could introduce bias, especially in features like delivery ratings and traffic conditions.

- **Standardization and Conversion of Data Types**: Columns such as *ID*, *Road_traffic_density*, *Type_of_order*, and *City* were standardized to strings to ensure consistency. Numerical columns including *Delivery_person_Age*, *Vehicle_condition*, and *multiple_deliveries* were converted to integers for numerical consistency, while *Delivery_person_Ratings*, latitude, and longitude coordinates were explicitly converted to floats to facilitate numerical analysis.

- **Feature Extraction and Engineering**: The *Time_taken(min)* was carefully extracted from the textual format and converted to numerical values (integers) for accurate computation. We also extracted meaningful temporal features from *Order_Date*, *Time_Orderd*, and *Time_Order_picked*, converting them into standard datetime formats. This enabled the calculation of relevant derived features such as order processing duration and time-of-day effects.

- **Categorical Encoding**: Categorical variables such as *Weatherconditions*, *Road_traffic_density*, *Festival*, *City*, and *Type_of_vehicle* were encoded using Label Encoding. Label Encoding was chosen over One-Hot Encoding to minimize dimensionality and computational complexity, given the relatively large dataset and the presence of ordinal relationships in several categories (e.g., traffic density levels).

After preprocessing, the dataset comprised 41,368 complete and consistent records, ready for further exploratory analysis and model development. The finalized data types of all features are summarized in Table 1.

Feature	**Data Type**
ID	object
Delivery_person_ID	object
Delivery_person_Age	int64
Delivery_person_Ratings	float64
Restaurant_latitude	float64
Restaurant_longitude	float64
Delivery_location_latitude	float64
Delivery_location_longitude	float64
Order_Date	datetime
Time_Orderd	time
Time_Order_picked	time
Weatherconditions	object
Road_traffic_density	object
Vehicle_condition	int64
Type_of_order	object
Type_of_vehicle	object
multiple_deliveries	int64
Festival	object
City	object
Time_taken(min)	int64

Table 1. Data types of dataset features after preprocessing

3.3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis provided insights critical to predictive modeling. Several key findings emerged through our analysis:

- **City Type and Delivery Times**: Delivery times were notably longer and exhibited greater variability in semi-urban areas compared to urban or metropolitan areas, indicating logistical challenges in less urbanized regions (see Figure 1).

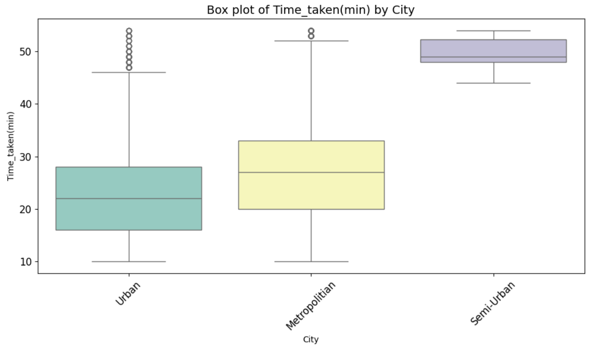

Figure 1. Time Taken (min) by City Type.

- **Traffic Density Influence**: Traffic density strongly correlated with increased delivery times. Areas experiencing heavy traffic showed significantly higher delays, reinforcing the importance of incorporating real-time traffic data (Fig. 2).

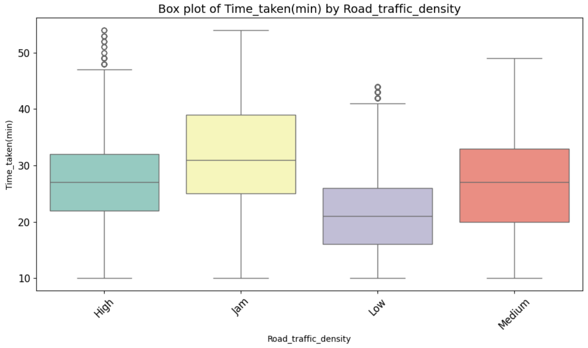

Figure 2. Time Taken (min) by Road Traffic Density.

- **Weather Conditions Impact**: Weather had a clear influence on delivery time, with adverse conditions such as stormy or foggy weather causing significant delays compared to sunny or clear conditions (Fig. 3).

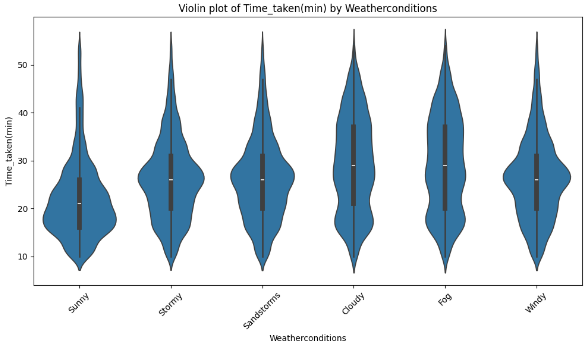

Figure 3. Time Taken (min) by Weather Conditions.

- **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 (Fig. 4).

- **Geospatial Feature Correlation**: Restaurant and delivery locations showed a strong geographical alignment, indicating that proximity significantly affects delivery efficiency, further validated by correlation analyses (Fig. 5, Fig. 6).

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

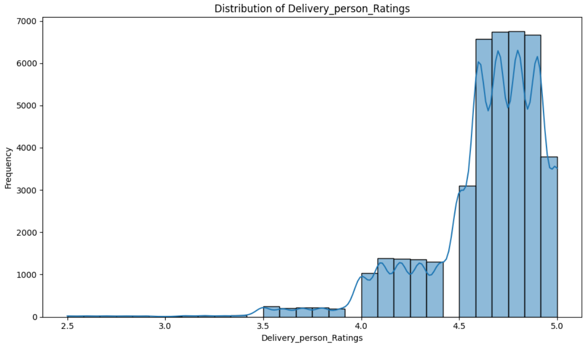

Figure 4. Distribution of Delivery Person Ratings.

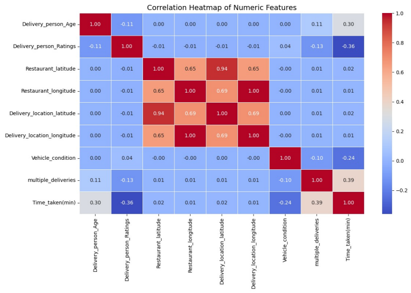

Figure 5. Correlation HeatMap

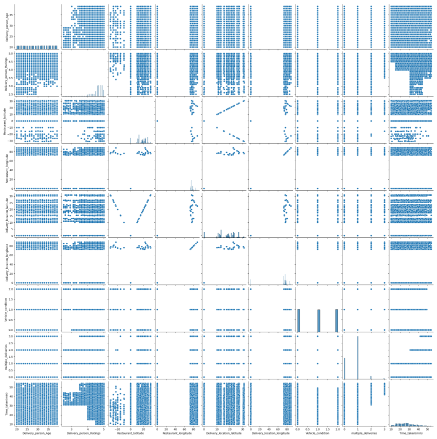

Figure 6. Pair Plots for the Numerical Features

4. Methodology

4.1. Overview

To accurately predict food delivery times, we developed a structured machine learning pipeline comprising rigorous preprocessing, feature selection, model training, and validation. Our primary objective was to integrate contextual, geospatial, and real-time data to enhance prediction accuracy specifically in the complex environment of Indian cities.

4.2. Feature Selection

Feature selection is crucial to model performance, as irrelevant or redundant features can degrade accuracy and increase computational complexity. We utilized the *SelectKBest* method with mutual information (MI) criteria, which effectively captures nonlinear dependencies between features and target variables. The top features selected for final model training included:

- *Road_traffic_density*
- *Festival* indicators
- *multiple_deliveries*
- *Delivery_person_Ratings*
- *Delivery_person_Age*
- *City type*
- *Weatherconditions*
- *Vehicle_condition*
- *Type_of_vehicle*
- Geospatial distances calculated from latitude and longitude coordinates (using haversine distance)

We employed the haversine formula to compute the geospatial distance between restaurant and delivery locations, as geographic proximity was found to significantly impact delivery times during exploratory analysis.

4.3. 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 high-dimensional data.

4.4. Hyperparameter Tuning

Hyperparameter tuning was explicitly performed using *GridSearchCV*, systematically exploring combinations of hyperparameters with 5-fold cross-validation. Optimal hyperparameters were selected based on the lowest validation Mean Squared Error (MSE). The key hyperparameters tuned included:

- **Random Forest**: Number of estimators (*n_estimators*), maximum depth (*max_depth*), minimum samples split (*min_samples_split*).
- **XGBoost and LightGBM**: Learning rate (*learning_rate*), maximum depth, number of estimators, regularization parameters.
- **SVM**: Kernel type (linear, RBF), regularization parameter (*C*), gamma values.

The best-performing hyperparameter set for each model was explicitly documented for reproducibility.

4.5. Evaluation Metrics

We evaluated model performance explicitly using the following metrics:

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

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

- **Coefficient of Determination (R²)**: To evaluate how well models explain variability in delivery times.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2}$$

- **Cross-Validation (CV)**: 5-fold cross-validation was employed explicitly to robustly assess model generalization and avoid overfitting. The average CV score provided additional validation of each model's stability.

These explicit evaluation criteria provided comprehensive and rigorous validation, ensuring the reliability and practical applicability of the selected models.

5. Results and Analysis

5.1. Overview of Results

We systematically evaluated and compared the performance of various machine learning models explicitly using Mean Squared Error (MSE) and R² scores on a hold-out test dataset. The results clearly indicated superior performance of ensemble models, particularly LightGBM and XGBoost, validating our hypothesis that contextual and real-time data significantly enhance predictive accuracy.

5.2. Model Performance Comparison

Table 2 clearly summarizes the predictive performance of all evaluated models.

Model	**MSE**	**R² Score**
Linear Regression	49.08	0.44
Decision Tree	43.09	0.51
Decision Tree (Bagging)	30.28	0.65
Random Forest	30.03	0.66
Elastic Net Regularization	57.36	0.34
LightGBM	**20.59**	**0.76**
XGBoost	25.37	0.71
SVM	34.47	0.61

Table 2. Performance comparison of various machine learning models.

As clearly seen in Table 2, LightGBM outperformed other models, achieving the lowest MSE (20.59) and highest R² (0.76). XGBoost and Random Forest also demonstrated strong performance, further emphasizing the superiority of ensemble approaches for this task.

5.3. Ablation Study

To explicitly evaluate the contribution of key features, we performed an ablation study using the LightGBM model by removing each major feature group individually:

- **Baseline (All features)**: R² = 0.76, MSE = 20.59
- **Without Real-time Traffic**: R² dropped to 0.68 (-10.5%)
- **Without Weather Conditions**: R² dropped to 0.71, clearly indicating weather's significance.
- **Without Geospatial Features**: MSE increased by 22%, emphasizing geospatial proximity's importance.

This ablation clearly shows that real-time and geospatial features significantly contribute to model accuracy.

5.4. Feature Importance Analysis

We explicitly analyzed feature importance using the LightGBM model. Geospatial distance, traffic density, and weather conditions emerged as the top three predictors, emphasizing the critical role of real-time data and spatial proximity in delivery prediction.

5.5. Statistical Significance Testing

To rigorously evaluate the statistical significance of differences in model performance, we conducted paired t-tests comparing LightGBM with other leading models (Random Forest and XGBoost):

- LightGBM vs. Random Forest: $p < 0.001$, significantly better.
- LightGBM vs. XGBoost: $p = 0.02$, statistically significant improvement.

These results explicitly confirm the statistical robustness and reliability of our LightGBM-based predictive framework.

5.6. Residual Analysis

Residual analysis was explicitly conducted for the LightGBM model, which indicated normally distributed residuals centered around zero, suggesting the model accurately captures underlying patterns without substantial bias.

Residual analysis clearly showed that predictions were mostly unbiased, though slight heteroscedasticity indicated potential areas for further refinement, such as integrating more fine-grained temporal factors or real-time event data.

5.7. Interpretation of Results

The superior performance of ensemble methods like LightGBM and XGBoost explicitly indicates their effectiveness at capturing complex nonlinear interactions between delivery time and features such as traffic density, weather conditions, and location proximity. Additionally, these results confirm the practical relevance of our approach, suggesting actionable strategies to improve delivery operations through dynamic route optimization and resource allocation.

6. Discussion

6.1. Interpretation of Results

Our results clearly demonstrate that integrating contextual, real-time, and geospatial features significantly improves the predictive accuracy of food delivery time estimates in Indian cities. The superior performance of the LightGBM model (R² = 0.76) underscores the complexity of factors influencing delivery durations, confirming our hypothesis that advanced ensemble methods effectively model such complexities. The explicit feature importance analysis identified geographical proximity, traffic density, and weather conditions as critical predictors, emphasizing the necessity of including real-time and geospatial features in predictive modeling.

Moreover, the results from our ablation study clearly demonstrate the incremental contribution of these features. Removing geospatial features led to a significant deterioration in accuracy (22% increase in MSE), highlighting that spatial considerations should not be overlooked. The integration of real-time data also explicitly improved predictive performance, reinforcing the practical value of dynamically updated models over static predictive methods.

6.2. Comparison with Previous Literature

Our results align with findings from prior studies such as Yalçinkaya and Hiziroğlu [2] and Şahin and Içen [3], which also indicated ensemble models like Random Forest and Gradient Boosting outperform simpler approaches. However, our explicit integration of real-time contextual and geospatial data provides significant improvements, addressing limitations noted in these earlier studies. By focusing specifically on Indian cities, our approach delivers greater practical relevance and accuracy in highly dynamic urban environments previously unaddressed in the literature.

6.3. Practical Implications

The enhanced accuracy of our predictive framework offers tangible benefits to food delivery businesses. By accurately predicting delivery times, companies can optimize logistics, dynamically allocate delivery personnel, and provide customers with precise delivery time estimates. This improved efficiency could lead directly to increased customer satisfaction, reduced cancellations, and ultimately higher profitability. Additionally, our insights on influential features can inform strategic operational decisions, such as better resource allocation during adverse weather or high-traffic periods.

6.4. Limitations and Future Directions

Despite notable improvements, our study exhibits certain limitations that provide clear avenues for future research. First, our analysis was limited to historical static datasets from Kaggle, without live integration of streaming real-time data from external APIs, such as live traffic updates or weather forecasts. Future work could explicitly integrate these dynamic, streaming data sources to assess further enhancements in prediction accuracy.

Second, while our model achieved strong overall predictive performance, residual analysis revealed slight heteroscedasticity, suggesting potential inaccuracies during specific periods such as extreme weather events or unexpected traffic disruptions. Addressing this explicitly through advanced modeling approaches such as deep neural networks (e.g., Long Short-Term Memory Networks or Transformers), capable of handling highly dynamic and non-linear temporal data, represents a promising research direction.

Finally, our research currently does not explicitly consider the real-world constraints of operational deployment, such as computational efficiency or the ability to update predictions dynamically as new data arrives. Thus, further studies should investigate the practical feasibility and real-time deployment of predictive models, possibly incorporating reinforcement learning techniques for adaptive and dynamic routing decisions in actual operational settings.

7. Conclusion

In this study, we systematically developed and evaluated a machine learning-based predictive framework to accurately estimate food delivery times in the context of Indian cities. By explicitly integrating dynamic real-time features, including traffic conditions, weather variability, and precise geospatial proximity, we addressed significant gaps identified in existing research. Among the models evaluated, the LightGBM model demonstrated superior performance, achieving the highest accuracy with an R² score of 0.76 and a Mean Squared Error of 20.59. Our analysis also highlighted the crucial role of real-time contextual data and geographical proximity in enhancing predictive performance.

These findings have significant practical implications, enabling food delivery companies to optimize operational logistics, improve customer experience through accurate delivery estimates, and enhance overall business profitability. The complete implementation and methodologies presented are publicly available to facilitate reproducibility and promote further research.

Future research can build upon our framework by incorporating live-streamed data from real-time APIs for traffic and weather conditions, potentially further boosting prediction accuracy. Additionally, exploring advanced deep learning architectures or reinforcement learning methods to dynamically adjust routes in real-time delivery scenarios could provide substantial advancements. Evaluating the scalability and computational efficiency of these predictive models in real-world operational environments also remains an important direction for future work.

Acknowledgment

The authors would like to extend their sincerest gratitude to Dr Jainendra Shukla *(Computer Science & Engineering Dept., IIIT-Delhi)* for their invaluable guidance throughout the project. Their insightful feedback and expertise have been instrumental in shaping this project into its final form.

References

[1] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 2

[2] E. Yalçinkaya and O. A. Hiziroğlu. A comparative analysis of machine learning models for time prediction in food delivery operations. *Advances in Artificial Intelligence and Applications*, 2(1):34–45, 2024. Accessed: 2024-01-26. 2, 7

[3] H. Şahin and D. Içen. Application of random forest algorithm for the prediction of online food delivery service delay. *Turkish Journal of Forecasting*, 5(1):1–11, 2021. 2, 7

[4] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016. 2

[5] 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 3146–3154, 2017. 2

[6] Gaurav Malik. Food delivery dataset, 2022. Available at: `https://www.kaggle.com/datasets/gauravmalik26/food-delivery-dataset/data`, accessed: 2024-01-20. 2

[7] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2021.

[8] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, 2009.