FOOD DELIVERY TIME PREDICTION

Siddharth Gavhane, Suyash Arlimar, Dhiraj Landge
University at Buffalo, School of Engineering and Applied Sciences, Industrial Engineering Program,
Buffalo, NY 14221, USA
gavhane@buffalo.edu suyashra@buffalo.edu dhirajsa@buffalo.edu

---

ABSTRACT

A predictive delivery time feature is presented for a food delivery application. As a result of an increasing demand for convenience and efficiency, this innovative feature uses data analytics and machine learning algorithms to estimate accurate delivery times for food orders. This system enhances the user experience by providing customers with precise delivery estimates based on various factors, including order volume, distance, traffic conditions, and kitchen preparation times. In the competitive food delivery industry, the importance of predictive delivery time for meeting customer expectations, improving operational efficiency for restaurants, and increasing overall user satisfaction is highlighted in this abstract.

This functionality also aids to operational efficiency while meeting consumer expectations. It may considerably enhance overall performance by allowing restaurants and delivery services to manage resources more effectively and reducing the likelihood of late deliveries or cold food. Furthermore, the system is based on data – driven insights, collecting and analyzing data on order history, traffic patterns, and kitchen performance on a constant basis to produce increasingly exact forecasts. Offering an adequate predictive delivery time feature could prove a key difference in the intensely competitive food delivery industry. It both attracts and maintains consumers who seek constant and trustworthy service.

Keywords: Predictive delivery time, data analysis, order volume, distance, traffic conditions, and kitchen preparation times, operational efficiency.

1. INTRODUCTION

In the fast changing food delivery market, this detailed study attempts to improve customer experience by developing a predicted delivery time feature for a meal delivery application. The study dives into calculating delivery times by taking into account many influencing aspects such as order volume, distance, traffic conditions, city kinds, weather conditions, and delivery crew features. The research uses data analytics and machine learning algorithms to enhance these predictions using a dataset from Zomato, an Indian food delivery service.

The technique incorporates ideas from well-known scholars in the field. It extends Zheng Wang, Kun Fu, and Jieping Ye's (2018) work on the important role of machine learning in travel time prediction. The study also improves on Chengliang Gao et al.'s (2021) FDNET deep learning solution, which is noted for its precision in time prediction. It also investigates route optimization tactics, incorporating the findings of Heng Tao Shen, Xiaofang Zhou, and Zaiben Chen (2011), as well as Alajali et al. (2018) and Richard Barnes et al. (2020), to enhance time estimations and route prediction models. Hanife AHN and Duygu EN (2021) emphasized the effectiveness of Random Forest for categorizing delivery times, while Elham Pourrahmani, Miguel Jaller, and Dillon T. Fitch-Polse (2023) provided insights on the relevance of delivery distance and wait times.

In addition, the the study brings from Zhu Xiaodi's (2022) exploration of machine learning models during COVID-19, Yanru Zhang and Ali Haghani's work on gradient boosting regression trees, and Hongrui Chu et al. (2023) and Aman Kharwal (2022) analyses of neural network models and last-mile delivery optimization. It also incorporates research on the importance of driver behavior and predictive modeling in delivery time prediction by Sheng Liu, Long He, and Zuo-Jun Max Shen (2020), Veridiana Rotondaro Pereira et al. (2023), and studies on customer satisfaction and service time prediction in delivery time estimates by Wenjie Wang and Li Jiang (2022) and Jie Zheng et al. (2022).

The main objective of this comprehensive study is to position the application as a trustworthy and efficient choice in the competitive marketplace, with the goal of increasing consumer satisfaction and operational efficiency for delivery services and restaurants. The document's methodology, data analysis, and conclusions—which elaborate on the results and the various effects of predictive delivery time features in the food delivery industry—will be covered in the parts that follow.

2. LITERATURE REVIEW

This literature review provides a comprehensive examination of research related to the prediction of delivery times within the food delivery services industry. Researchers examine the application of machine learning and data-driven techniques, including deep learning models, Random Forest, and gradient boosting regression, to enhance delivery time estimates to enhance customer satisfaction and operational efficiency. Furthermore, some studies emphasize the inclusion of factors such as service time and driver behavior in prediction models, while others emphasize the optimization of meal delivery routes. These papers demonstrate the importance of precise time predictions in the food delivery industry, offering valuable insights for improving the user experience and advancing the field.

Zheng Wang, Kun Fu, Jieping Ye (August 19-23, 2018), this paper delves into the critical realm of travel time estimation and its application to food delivery services, with a focus on overcoming existing limitations. It explores traditional methods for travel time estimation, highlighting their shortcomings, and examine the role of machine learning techniques, including regression algorithms, in improving accuracy. Furthermore, this discusses the deep learning model proposed in the research, comprising wide, deep, and recurrent components adept at handling complex feature representations. The review underscores the significance of location-based data and feature engineering in the context of food delivery. This paper also present findings from both offline evaluations with extensive vehicle travel data and real-time assessments on the DiDi platform, showcasing the model's superior performance. In conclusion, this paper not only emphasizes the importance of precise travel time estimation in food delivery apps but also outlines promising future directions, such as feature system enrichment and the applicability of the proposed deep learning model to various location-based regression problems.

Chengliang Gao, Fan Zhang, Guanqun Wu, Qiwan Hu, Qiang Ru, Jinghua Hao, Renqing He, Zhizhao Sun (August 14–18, 2021), in the context of predicting delivery times for a food delivery app, the study presents FDNET as a deep learning solution that substantially enhances the accuracy and efficiency of time prediction in online food delivery services. The empirical evidence showcased in the paper highlights FDNET's superiority compared to conventional models, which is a valuable insight for improving the reliability of delivery time estimates. Furthermore, the paper's emphasis on future research directions, particularly personalization and real-time data integration, offers valuable guidance for enhancing the performance of time prediction algorithms. As the accuracy of delivery time estimates is a pivotal factor in enhancing customer satisfaction and optimizing driver experience in food delivery apps, the findings, and methodologies of FDNET hold great promise for advancing the state of the art in this field.

Zaiben Chen, Heng Tao Shen, Xiaofang Zhou (2011), while the paper discussed the problem of identifying popular routes for transportation, its methodologies and findings can be adapted and applied to the context of predicting delivery times for food delivery apps. Just as the Coherence Expanding algorithm was used to extract transfer networks from travel trajectories, it can be employed to extract insights from historical delivery data, such as driver routes and order fulfillment times. The robust popularity indicator used for transfer nodes can be repurposed to assess factors that affect delivery times, such as restaurant popularity, traffic conditions, and order volume. Leveraging these insights, the Maximum Probability Product algorithm can be adapted to predict estimated delivery times, considering popularity and other relevant factors. While the focus of this literature review is on route discovery, it demonstrates the potential for applying similar data-driven techniques to optimize and predict delivery times in food delivery apps, offering valuable insights for the food delivery industry and enhancing the user experience. Further research and refinement in this direction are warranted to leverage the promising outcomes highlighted in the reviewed paper for the specific context of food delivery services.

Alajali, W., Zhou, W., Wen, S., & Wang, Y. (2018), this paper delves into the booming urban food delivery sector, emphasizing the increasing importance of precise route predictions, especially in highly populated regions. Their research predominantly centers on immediate route forecasting models, drawing a distinction between non-parametric techniques, which recognize intricate non-linear urban paths, and parametric strategies like ARIMA and SVR. Notably, while the kNN emerges from the non-parametric cluster, its challenges with scalability are underscored. The paper comments on the significant computational demands of neural models such as RBF and deep learning but also concedes their potential precision. The simplicity of decision trees is recognized, tracing their evolution into robust ensemble methodologies like GBRT, RF, and XGBoost. Additionally, the FIMT-DD strategy's relevance in online learning is spotlighted, showcasing its adaptability. Contrarily, Smith and his team emphasize the urgent need to rigorously evaluate data protection and individual privacy in this rapidly advancing domain.

Richard Barnes, Senaka Buthpitiya, James Cook, Alex Fabrikant, Andrew Tomkins, Fangzhou Xu (August 23–27, 2020), this paper begins by introducing the significance of accurate delivery time predictions in food delivery apps and highlight the relevance of our research objective, inspired by the success of BusTr in public transit prediction. This paper provides an overview of the BusTr model's key features, and its superior performance compared to previous models, including DeepTTE. This paper discusses the challenges specific to food delivery time prediction, emphasizing the importance of feature selection and design choices, as well as the need for sufficient training data. Furthermore, this paper underscores the potential benefits of reducing uncertainty in delivery times for food delivery services and suggest future directions for research in this area. Throughout the review, this paper emphasizes the role of the work in advancing the field and reducing the uncertainty associated with transit times, ultimately enhancing user experience and efficiency in food delivery apps.

Elham Pourrahmani, Miguel Jaller, Dillon T. Fitch-Polse (2023), the delivery distance is significant factor for delivery wait time irrespective of delivery fee, this includes traffic congestion and consumer density which could be different for different location in city. Wait time for midday deliveries are lower compared to earlier times of time, due to high supply and demand for food deliveries. Wait time for DoorDash is lower compared to other apps due to established network and registered customers.

Hanife ŞAHİN, Duygu İÇEN (August 2021), there are several benefits to using Random Forest (RF) to categorize food deliveries as "on time" or "late" when forecasting delivery times for meal delivery applications. It is especially useful because RF can effectively consider a wide range of factors affecting delivery delays. RF produces more accurate and trustworthy time forecasts by reducing variation and preventing over fitting. Accurate projections across a range of situations depend on its capacity to comprehend complex, nonlinear interactions among variables. Additionally, by using RF's insights into key predictors to optimize delivery processes, the field of food delivery services will see higher consumer happiness and increased operational efficiency.

Yanru Zhang, Ali Haghani, to improve the precision and understandability of delivery time forecasts, food delivery applications can use gradient boosting regression trees. Like its effect on improving highway traffic time forecasts, this method can improve the estimation of food delivery times by accommodating a wide range of predictor variables, addressing complex relationships, fine-tuning model parameters, and facilitating comparisons with alternative models. This guarantees that the projected delivery times given to customers via mobile applications are accurate and understandable, enhancing user happiness and operational effectiveness within the food delivery service industry.

Zhu, Xiaodi (Dec 2022), this paper examines the potential benefits of machine learning models in examining the factors that influence the popularity of food delivery apps during the COVID-19 pandemic. It highlights the importance of achieving high accuracy in predicting delivery times, as demonstrated by models such as CatBoost, and the need to understand user behavior patterns to accurately predict delivery times. Additionally, it highlights the potential of deep learning to model complex relationships that affect delivery times, as well as the influence of compatibility, online reviews, and other factors known to influence app popularity. Ultimately, the paper concludes that the use of machine learning methods to improve delivery time predictions can lead to increased user satisfaction and improved operational efficiency in the food delivery industry.

Aman Kharwal (2022), a sequential neural network model, combining LSTM with DENSE layers, is employed to predict the time needed for food delivery, considering variables such as distance, travel time of the driver, traffic, weather, age of the driver, vehicle type, and vehicle condition. This approach facilitates the modeling of complex relationships between variables and offers real-time adaptation to changing conditions, thus improving user satisfaction by providing accurate and reliable delivery time estimates. Additionally, it optimizes operational efficiency by helping to allocate resources and plan routes, thus responding to the ever-changing demands of the food service industry and enhancing the overall user experience.

Hongrui Chu, Wensi Zhang, Pengfei Bai, and Yahong Chen (2023), this paper tackles the intricate challenges and explore potential remedies related to enhancing last-mile delivery in online food service platforms. Acknowledging the pivotal role of punctual delivery, their investigation identifies delivery time as a crucial yet unpredictable element in both order assignment and route planning. The collective researchers introduce a groundbreaking, data-fueled optimization method, accentuating the intrinsic difficulties in accurately predicting variables like driver actions and traffic conditions. This innovative strategy melds machine learning methodologies with capacitated vehicle routing optimization in a cooperative fashion. The researchers endorse a refined predict-then-optimize (SPO) framework, offering it as a successor to the conventional "predict-then-optimize" model. Here, choice error, instead of prediction error, forms the prediction objective, facilitating more accurate optimization outcomes. The research additionally embraces multi-faceted data, spanning from meteorological and seasonal variances to instantaneous traffic information, underscoring its immense potential to significantly enhance decision-making in the brisk environment of online food delivery. Their results indicate that the SPO method surpasses preceding methods by approximately 5%, underscoring its relevance in the contemporary food delivery landscape.

Jie Zheng, Ling Wang, Shengyao Wang, Jing-fang Chen, Xing Wang, Haining Duan, Yile Liang, Xuetao Ding (2022), this paper intensively studied the intricacies of time-related uncertainties in the on-demand food delivery (OFD) sector, emphasizing the vital role of service time. Although crucial for delivery projections, service time is often neglected in existing research. These scholars introduced a novel approach to forecast the volatile nature of service time by applying a Gaussian mixture model (GMM). They identified a considerable void in contemporary studies that primarily concentrate on predicting travel and food preparation durations but sideline service time. To address this, they reconceptualize the distribution estimation challenge as a clustering issue, adopting an innovative hybrid estimate of distribution algorithm (HEDA) for solutions. Evaluations on the Meituan platform, through both offline and online assessments, confirm the consistency and accuracy of their techniques in enhancing time prediction precision for OFD services.

Wenjie Wang, Li Jiang (7 July 2022), this paper explores the crucial realm of time prediction for online food delivery apps, with a focus on enhancing customer satisfaction in the context of heterogeneous time-sensitive customers. It begins by underlining the pivotal role of accurate time prediction in online food delivery, emphasizing its direct impact on customer satisfaction and overall platform success. The paper delves into the existing body of literature concerning optimization techniques employed in meal delivery, including route optimization and clustering, while stressing the trade-offs between cost minimization and customer satisfaction maximization. It offers an in-depth examination of the proposed two-stage solution involving Hierarchical Agglomerative Clustering (HAC) and Genetic Algorithms (GA) for optimizing meal delivery routes to meet customer time satisfaction, highlighting their effectiveness in delivering meals within specified time windows. Additionally, the study underscores the importance of factoring in customer time sensitivity and suggests that meal delivery platforms should establish time-sensitive customer profiles to inform routing decisions. It concludes by outlining potential avenues for future research, such as accounting for traffic conditions, real-time demand fluctuations, and integrating multiple objectives into optimization strategies, ultimately highlighting the need for adaptive algorithms in dynamic delivery scenarios.

Veridiana Rotondaro Pereira, Ana Maria Saut, Asta Carolina Dogas Pieralini, Orlando Yesid Esparza Albarracin (July 2023), this paper is focused on enhancing the operational efficiency of an online food delivery app by monitoring and predicting food delivery lead times, this paper employed Shewhart control charts to monitor mean delivery lead times using historical data. To account for inherent variability, this paper established dynamic control limits that change based on the day of the week. Additionally, it identified seasonal trends in delivery lead times, with increases during weekdays and a drop during weekends, possibly linked to changing working conditions during the pandemic period. We've underscored the significance of efficient food delivery in ensuring food safety and noted the limitations of convenience sampling approach. The ongoing efforts involve predictive modeling, real-time monitoring, and optimization strategies to improve delivery efficiency, with a commitment to data-driven decision-making and continuous improvement.

Sheng Liu, Long He, Zuo-Jun Max Shen (2020), this paper explored the challenges faced by food delivery services in guaranteeing punctual last-mile deliveries. Partnering with a Chinese delivery firm, their research highlighted the significant role of drivers' route choices and the unpredictability of service durations in delivery tasks. These aspects, often underrepresented in conventional systems, greatly affect delivery punctuality. Their work introduces an innovative model, merging travel-time predictions with order assignment techniques, to better understand the often-hidden decision-making of drivers. This fusion helps in decoding drivers' route preferences, a task that's hard to quantify. This paper also found that drivers, even with advanced tools, tend to diverge from preset routes, relying more on their judgment and current situations. The researchers devised a practical method to predict total travel durations, bypassing the need for intricate route designs. Initial results using actual delivery data showed notable improvements in punctuality using their method, underscoring the need to factor in driver behaviors. The study ends suggesting further exploration into fixed order assignments, accounting for uncertainties in customer location and service durations.

Chun-Hsin Wu, Jan-Ming Ho, and D. T. Lee (2004), this paper explored the potential of using Support Vector Regression (SVR) to predict travel durations in transportation networks. The study emphasized the crucial role of accurate travel-time predictions in advancing intelligent transportation systems. Based on genuine highway traffic data, the findings leaned towards SVR due to its notable efficiency in minimizing relative mean errors and root-mean-squared errors. Citing Vapnik's foundational work, the research highlighted SVM's superiority over traditional neural networks, especially its unique ability to secure global minima. The study, in highlighting SVM's wide-ranging adaptability, also underscored the importance and potential of SVR in traffic data analysis.

3. Data collection and Description

Food delivery involves delivering meals or groceries from a restaurant, or store to a customer's home. Typically, customers place their orders online, through a mobile application, or through a dedicated food ordering platform. Orders may include main courses, side dishes, beverages, desserts, or basic grocery products packaged in boxes or bags. While delivery personnel often use cars for transportation, in densely populated urban areas, bicycles or motorized scooters might be more common. The dataset contains information about deliveries, related to a food delivery service. Based on this dataset, we can predict the time it will take to deliver the food to the customer after the order is placed. Each record appears to represent a unique delivery event which will help to predict the time. The dataset includes the following attributes:

ID:	A unique identifier for each record.
Delivery_person_ID:	An identifier for each delivery person.
Delivery_person_Age:	The age of the delivery personnel.
Delivery_person_Ratings:	A rating score, reflecting the performance of the delivery person.
Restaurant_latitude & **Restaurant_longitude:**	Geographic coordinates of the restaurant.
Delivery_location_latitude & **Delivery_location_longitude:**	Geographic coordinates of the delivery destination.
Order_Date:	The date when the order is placed.

Time_Ordered:	The time when the order is made.
Time_order_picked:	The time when the delivery person picked order.
Weather_conditions:	Indicates condition of weather at the time delivery.
Road_traffic_density:	An assessment of traffic conditions, ranging from 'Low' to 'Jam'.
Vehicle_condition:	A metric denoting the condition of the delivery vehicle.
Type_of_order:	The category of order, such as Snack, Meal, or Drinks.
Type_of_vehicle:	The kind of vehicle used for the delivery, such as motorcycle or car.
multiple_deliveries:	It Indicates if the delivery person is handling multiple orders simultaneously.
Festival:	Notes whether the delivery coincided with a festival.
City:	Specifies the type of city such as Metropolitan, Suburban, or urban.
Time_taken (min):	The duration for food delivery.
Time:	Difference between time ordered and time ordered picked.
Distance:	The distance covered for the delivery.

3.1 Numeric variables:

A numerical variable, also known as a quantitative variable, represents measurable quantities as numbers. From this dataset, the following variables are numeric variables.

ID
Delivery_person_ID
Delivery_person_Age
Delivery_person_Ratings
Restaurant_latitude & Restaurant_longitude
Delivery_location_latitude & Delivery_location_longitude
Order_Date
multiple_deliveries
Time_taken (min)
Distance

3.2 Categorical variables:

Categorical variables, also known as qualitative variables, represent data that can be categorized into distinct groups or categories without having an inherent numerical value. The following variable are categorical variables.

Weather_conditions
Road_traffic_density
Vehicle_condition
Type_of_order
Type_of_vehicle
multiple_deliveries
Festival
City

3.3 Timestamp Variables:

Timestamp variables are a type of data that record specific instances in time. The following variables are timestamp variables.

Time_Ordered
Time_order_picked

3.4 Response Variable:

Food delivery services rely on accurate delivery time predictions for both customer satisfaction and operational efficiency. This prediction is based on the response variable 'Time taken in minutes'. This variable represents the duration from the moment an order is placed at a restaurant to the time it is delivered to the customer. This variable is predicted based on several factors, including preparation time at the restaurant, distance from the restaurant to the customer's location, traffic conditions, mode of transportation, and delivery personnel efficiency.

4. DATA VISUALIZATION

In this exploratory data analysis, we analyze the dataset for food delivery, and we use data visualization as an attempt to understand some relations between predictors and response variables.

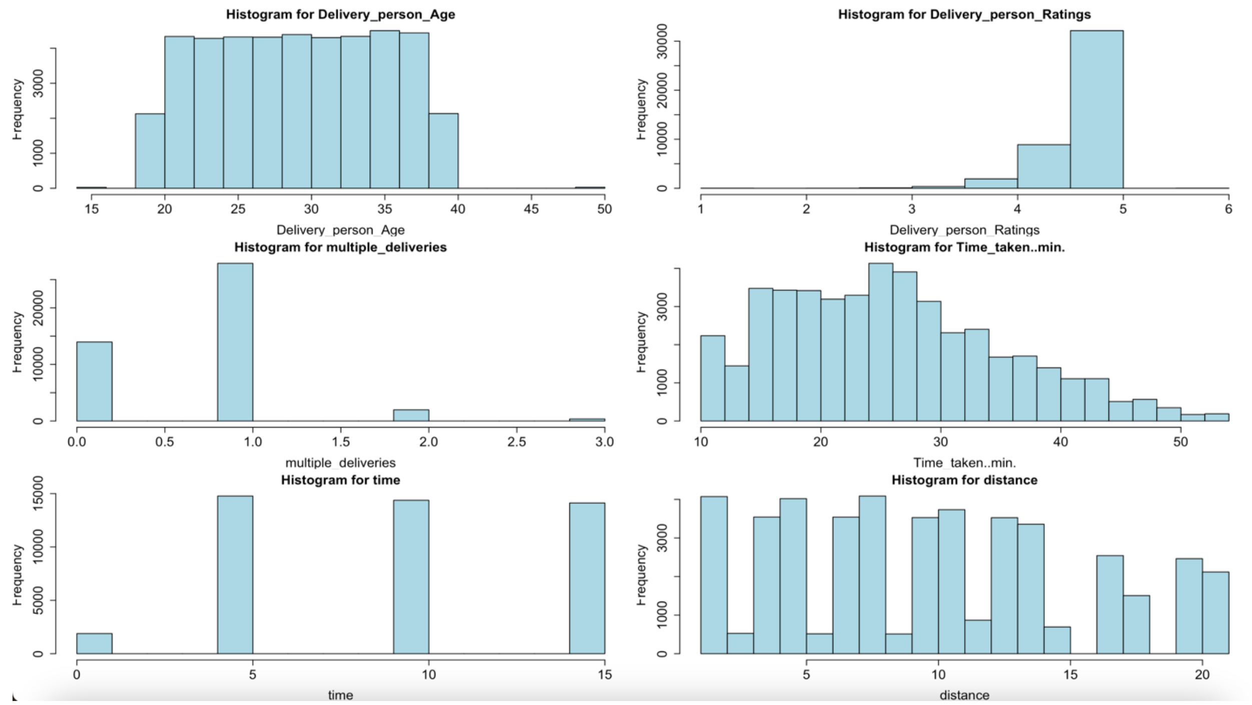

FIGURE NO 1: HISTOGRAM FOR NUMERICAL VARIABLES

These are the histograms of delivery person age, delivery person ratings, multiple deliveries, time taken in min, time, and distance from the original dataset. We get overall idea of the variables from these histograms which will help to brainstorm the methods to replace the NA values in the dataset.

During our data examination, we encountered missing values labeled as 'NA' and blank entries. Addressing these discrepancies was essential to ensuring the datasets reliability and facilitating accurate visualization and further analysis. We adopted Different methods, such as median and mode. The choice of imputation technique was steered by the data's characteristics, aiming to minimize potential biases.

Upon meticulous examination of our dataset, we identified that a total of seven columns are afflicted with missing values. The columns which have missing values are Delivery person Age, Delivery person Ratings, and multiple deliveries, weather conditions, road traffic density, festival, and city.

For doing further analysis it is important to consider the correlation between the independent variables and response variable.

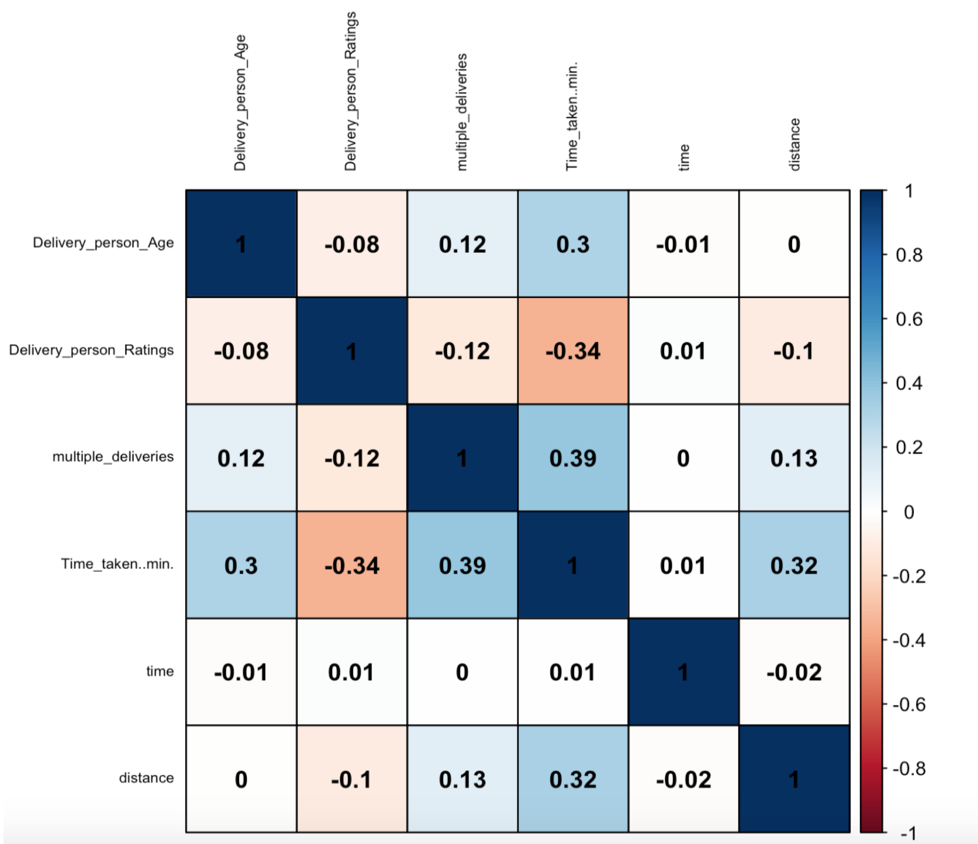

FIGURE NO 2: CORRELATION PLOT RESPONSE WITH INDEPENDENT VARIABLES

We conducted a correlation analysis on our delivery dataset to decipher the relationships between response variable, which is Time taken in min and independent variables, including the delivery person's age, ratings, multiple deliveries handled, time, and distance covered. Utilizing correlation coefficients, we quantified these relationships, with values close to 1 or -1 indicating strong correlations and those near 0 indicating weak ones.

The correlation coefficients of variables with respect to response variable are as follows:
Time taken in min & Delivery person age = 0.3
Time taken in min & Delivery person ratings = -0.34
Time taken in min & Multiple deliveries = 0.39
Time taken in min & time = 0.01
Time taken in min & distance= 0.32

The time taken for deliveries was significantly correlated with a delivery person's ratings by -0.34. Therefore, higher-rated delivery persons tend to complete deliveries faster. Also, a positive correlation of 0.39 between Multiple deliveries and Time taken in minutes indicates that the time taken in minutes increases as multiple deliveries increase, and the same positive correlation implies the age, time, and distance of the delivery person with the time taken in minutes. A correlation plot displaying positive correlations in blue and negative correlations in red provided immediate insight into the data.

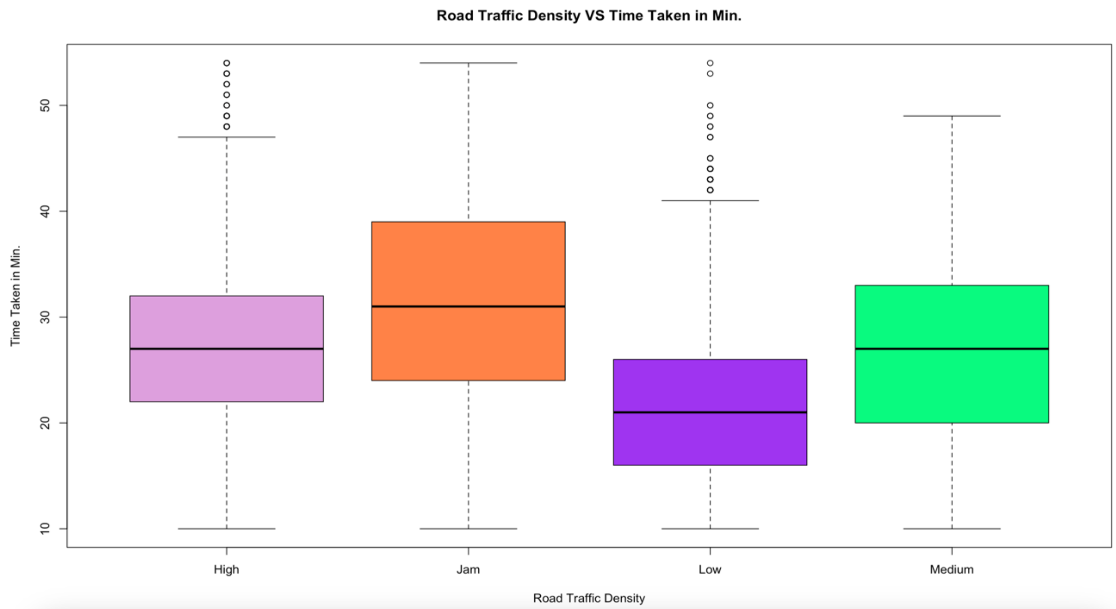

FIGURE NO 3: BOXPLOT FOR ROAD TRAFFIC DENSITY VARIABLE

Here we plotted boxplot for Road traffic density vs Time taken Min graph. This boxplot shows how road traffic density affects Time taken for delivery when traffic conditions are high, low, medium and jam. When traffic is high time taken for delivery is more compared to when traffic is low.

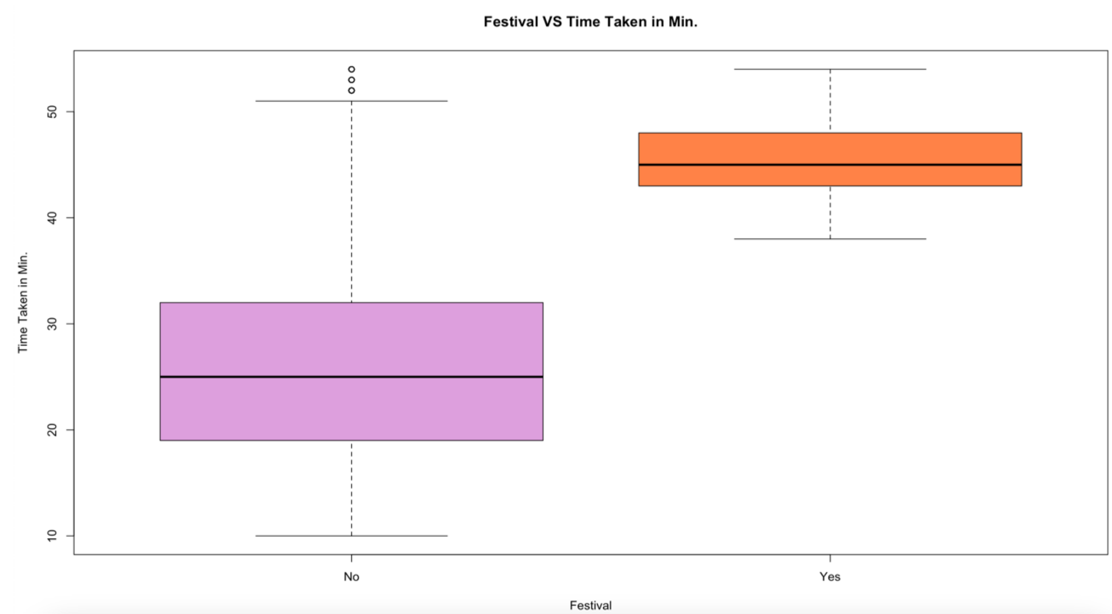

FIGURE NO 4: BOXPLOT FOR FESTIVAL VARIABLE

Here we plotted boxplot for festival variable and Time taken in Min variable (which is dependable variable). From graph we can see that time taken for delivery is more when there is festival.

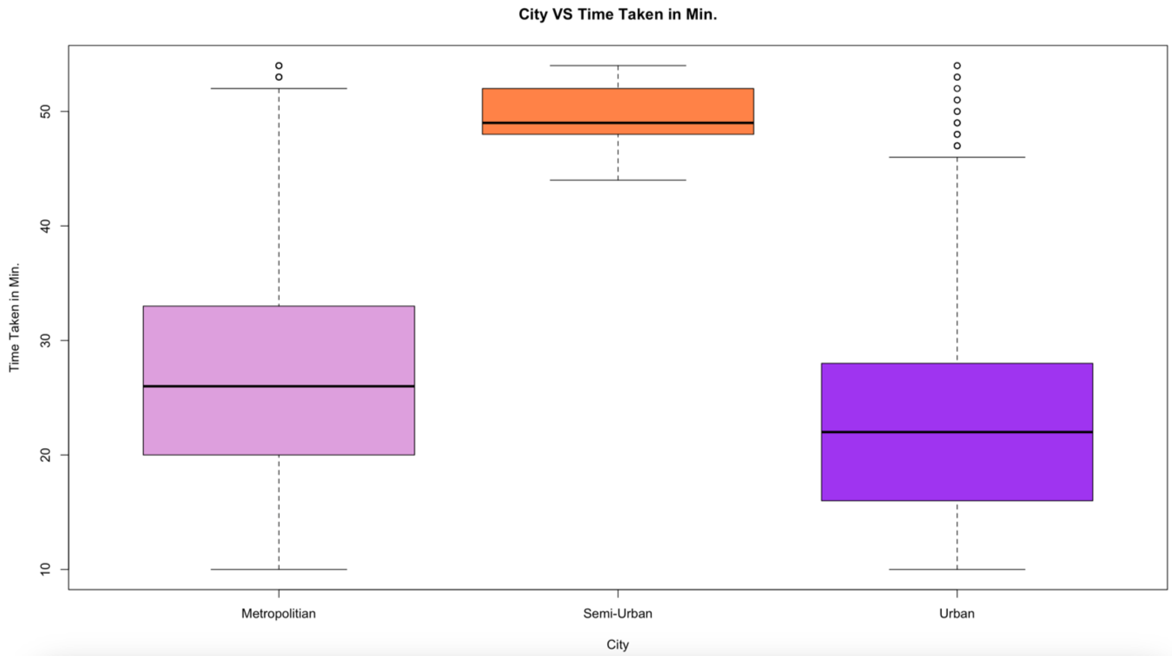

FIGURE NO 5: BOXPLOT FOR CITY VARIABLE

Here we plotted graph for city column and Time taken in min column. From graph we can see that, Time taken for semi urban is more compared to metropolitan and urban.

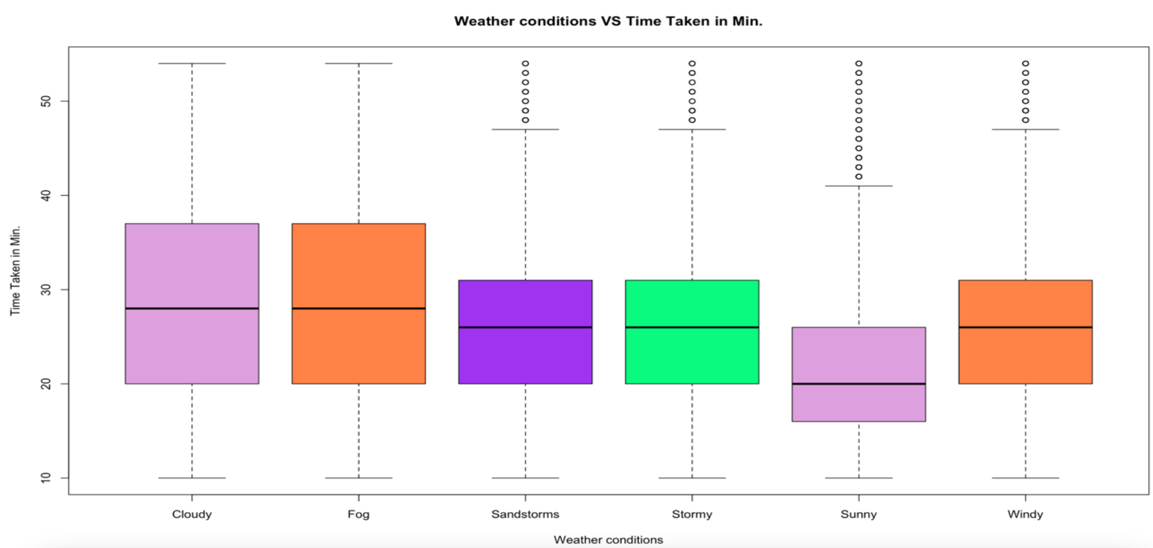

FIGURE NO 6: BOXPLOT FOR WEATHER CONDITIONS VARIABLE

Here we plotted boxplot for weather conditions and Time taken in Minutes. Graph shows time taken for delivery in different weather conditions.

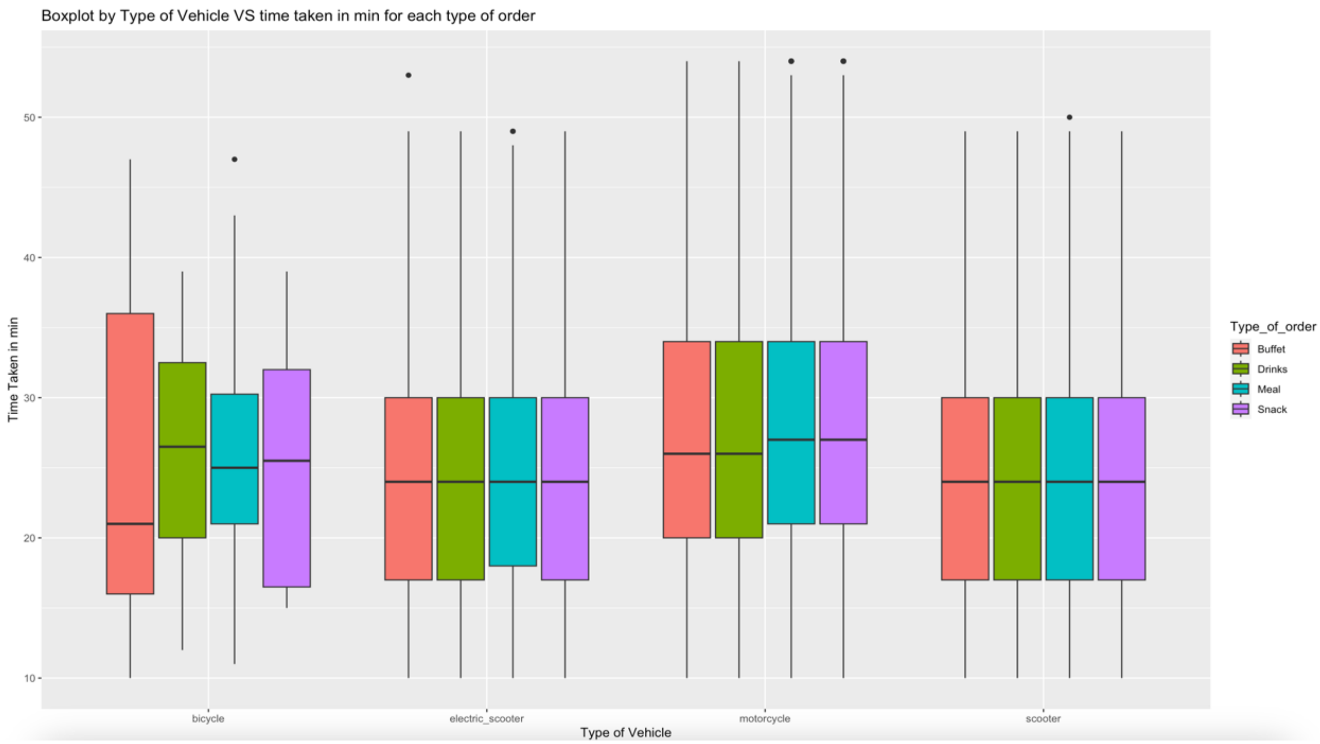

FIGURE NO 7: BOXPLOT FOR TYPE OF VEHICLE VARIABLE

Based on the boxplots, motorcycles are the most efficient vehicle type for deliveries, both in terms of speed and consistency. The delivery times for 'snack' orders are similar across all vehicle types, but for 'meal' orders, 'cars' and 'bicycles' lag. There are outliers in the 'bicycle' boxplot for certain order types that warrant further investigation.

5. METHODOLOGY

In our study, we execute a sequence of procedures for preparing and analyzing the dataset. This is done to predict the food delivery time.

5.1 Data Preprocessing:

In histogram, the Delivery person age column showcases a relatively uniform distribution of ages from around 20 to 40, with fewer delivery persons at the extremities of age 15 and age 50. As age is continuous variable and because of this data structure, median method is a best choice for addressing missing values. The median is robust to outliers and will not distort the distribution's central tendency.

We analyzed the histogram for delivery ratings. The graph indicated that a significant number of our delivery people have top-notch ratings, with many hovering near the score of 5. There's a smaller number with lower ratings. When faced with missing rating data, we chose to use the median. An average might get influenced by the occasional low scores, potentially giving a skewed view. The histogram supports this, showcasing a concentration of scores close to 5. Using the median ensures our data accurately reflects the general trend.

The chart shows that most delivery persons usually carry only one delivery at a time. Some don't carry any, and very few handle two or more at once. Because the most common number of deliveries is one, when we had missing data, we filled it in with mode of the data. This makes our data feel more genuine and accurate for further study.

Some of our data, like Road traffic density, Festival, and City, are categorical variables therefore we use a method that fills in the missing spots with the mode of the column. This keeps our data consistent and accurate. After filling in these gaps, we double-checked to make sure the changes were appropriate and didn't make the data biased. This ensures our data is reliable for future studies.

Also, there are some anomalies in the dataset that we need to fix like in person delivery ratings column, we have some ratings value more than 5 so it needs to be addressed to get quality data and accurate prediction.

We have coordinates of 2 different location, restaurant, and delivery. From latitude and longitude coordinates of these two locations we calculated distance between restaurant and delivery location using geosphere library. Firstly, defined R = 6371 which is radius of earth, then calculated distance using 'distcalculate' function as shown in code. Finally added distance column to data, showing distance between restaurant and delivery location.

5.2 Data Modeling:

Supervised learning models were employed to estimate the time required for food delivery. This approach involves using a collection of input variables to predict the time taken, measured in minutes. The study utilized a dataset comprising 22 variables, and a best subset model was applied to identify the most significant variables which are Festival, Delivery person ratings, distance, City semiurban and City urban. The dataset was divided into two parts, with 70% allocated for training and 30% for testing.

5.2.1 Linear Regression Analysis for Food Delivery Time Prediction

Multiple models were applied to the dataset to compare and check the accuracy, starting with Linear Regression, which served as the baseline model. In this model, the delivery time is predicted as a linear combination of the input features, providing an easily interpretable approach to understand the influence of each variable on the outcome. The effectiveness of the Linear Regression model is assessed using metrics such as R-squared and Root Mean Squared Error (RMSE). R-squared quantifies the percentage of variance in the dependent variable (Time taken in minutes) that can be explained by the independent variables, while RMSE calculates the average difference between the predicted values and the actual observations.

5.2.2 Ridge Regression Analysis for Food Delivery Time Prediction

A Ridge Regression technique has been implemented to mitigate overfitting and boost model performance. Using the 'glmnet' package in R, this method applies L2 regularization to constrain overly large coefficients, thus enhancing the model's ability to generalize and enhance accuracy on unseen data. The optimal regularization parameter (λ) is determined through cross-validation, which is a technique for assessing the model's generalizability. To meet the input requirements of the glmnet function, our training and test data are formatted into matrices. We train the model using the training data and then employ the optimally determined λ for making predictions on the test data. Root Mean Square Error (RMSE) measures how well the model performs by measuring the discrepancy between the predicted and actual values. Lower RMSE indicates better model performance. Additionally, we calculate the R-squared value to determine how much variance in the 'Time taken in minutes' is explained by the independent variables, with a higher R-squared suggesting a more effective model.

5.2.3 Lasso Regression Analysis for Food Delivery Time Prediction

This is a variant of linear regression that uses shrinkage to pull data points towards a central point, like the mean. It encourages the development of simpler, sparser models. To implement the lasso model in R, we use the 'cv.glmnet' function, setting alpha to 1 to ensure only the lasso penalty is used. This function also performs cross-validation to determine the best λ value, which is associated with the model’s highest level of predictive accuracy. The model is trained on our training dataset, and then we use the predict function with the optimal λ value to generate predictions on the test dataset. Lasso regression excels at diminishing the influence of less significant variables by reducing their coefficients to zero. We assess the model's performance through the RMSE, which quantifies the average discrepancy between predicted and actual delivery times. In addition, the R-squared value is calculated to measure the proportion of variance explained. The RMSE derived from our analysis offers an insight into the average deviation between predicted and observed delivery times.

5.2.4 Random Forest Analysis for Food Delivery Time Prediction

Random Forest is an ensemble learning method used in regression, which constructs numerous decision trees during training and outputs their average predictions. It is well known for its ability to handle large datasets with multiple variables, especially with respect to machine learning. Random Forest reduces the risk of overfitting by averaging the results of multiple trees, which may vary significantly on their own, to produce more accurate and stable predictions. Our Random Forest model was constructed using the ‘randomForest’ package in R, with 'Time_taken..min' designated as the response variable, and all other columns as predictors. Training was conducted on the entire training dataset, and predictions were generated using the predict function for the test dataset. RMSE is used to measure the accuracy of our Random Forest model in predicting delivery times. To gauge the amount of variance explained by the predictors, we calculate the R-squared value. As a result of its ability to handle nonlinear data and complex interactions between variables, this model is particularly good at predicting food delivery times. In addition to enhancing accuracy, its ensemble approach ensures consistency, making it less susceptible to outliers, by averaging multiple decision trees' predictions. Our high precision in estimating delivery times confirms the effectiveness and reliability of the Random Forest Regression model in various datasets.

5.2.5 Decision Tree Regression Analysis for Food Delivery Time Prediction

We used the Decision Tree Regression model, a straightforward but efficient method that divides the dataset into homogeneous subsets depending on feature values, with the 'rpart' package in R. The main benefit of this model is that it is simple to understand and visualize, and it offers clarity to decision-making processes. The model generated an RMSE, or average difference in projected delivery times, for our project, which was to estimate the timeframes for food delivery. The estimated R-squared value also shows the extent to which the model can account for the difference in delivery timeframes. These indicators imply that even while the Decision Tree might not fully reflect the dataset's intricacies, it is still a useful starting tool that provides crucial insights into the variables influencing delivery timeframes.

5.2.6 Gradient Boosting Regression Analysis for Food Delivery Time Prediction

We used the R 'gbm' package to implement the Gradient Boosting Regression approach, a sophisticated machine learning technique that improves prediction by gradually improving weaker models. It gradually accumulates up to 500 trees, each of which is intended to make up for the errors of the previous one. This method significantly improves the model's capacity to identify intricate nonlinear patterns within the dataset. We used R-squared and RMSE as performance metrics to evaluate the efficacy of our Gradient Boosting model. The R-squared value shows the percentage of data variance that the model accounts for, whereas the RMSE value aids in estimating the average error in the model's predictions. The model's ability to anticipate food delivery times accurately is demonstrated by the results of these measures, which also demonstrate the model's efficiency and excellent prediction precision in our analysis.

6. RESULTS AND DISCUSSION

In our project, we attempt to create a more effective delivery time prediction function for a food delivery app, addressing the sector's rising need for quick and efficient delivery. To anticipate delivery time periods, we used the power of machine learning and data analysis. Order size, distance, traffic, time, and the delivery persons' age and ratings were all factored into the model.

The research we conducted included a thorough examination of multiple machine learning models for their ability to reliably estimate delivery times. The precision and capacity to streamline the complexity of the models were evaluated. The linear regression model performed well, with an RMSE of 5.9325 and an R-squared of 0.6017. Ridge regression performed admirably as well, with an RMSE of 1.0854 and an R-squared of 0.9867. Lasso Regression, which is noted for its feature selection capacity, had an RMSE of 0.2740 and an R-squared of 0.9915. The Random Forest model, which excels at dealing with large datasets, produced an RMSE of 3.7425 and an R-squared of 0.8415. Notably, the Gradient Boosting model performed quite well in terms of accuracy, with an RMSE of 0.1146 and an R-squared of 0.9925, demonstrating its durability and efficacy in forecasting delivery time periods. These findings underlined the advantages of each algorithm and the significance of selecting the proper model for predictive task in the quick-paced food delivery field.

SR.NO.	MODEL NAME	RMSE value	R-squared value
1	Linear Regression	5.9325	0.6017
2	Lasso Regression	1.0854	0.9867
3	Ridge Regression	0.2740	0.9915
4	Random Forest	5.8312	0.6152
5	Decision Tree	3.7425	0.8315
6	Gradient Boosting	0.1146	0.9925

The predictive function improves operational efficiency and customer satisfaction significantly. It enables restaurants and delivery businesses to better manage their resources, decreasing incidences of late deliveries and cold food. Furthermore, continuous data gathering and analysis on order patterns, traffic, and kitchen performance update the system's forecasts, assuring increased accuracy and dependability.

The ability to give precise projected delivery times has become a critical distinction in the competitive world of food delivery. It not only appeals to clients looking for dependable service, but it also develops loyalty, which improves the brand's reputation. The project's future aims include further refining these algorithms and using real-time data and consumer input to improve forecast precision. However, maintaining accuracy will be difficult, especially given unpredictable external circumstances like as weather and traffic problems.

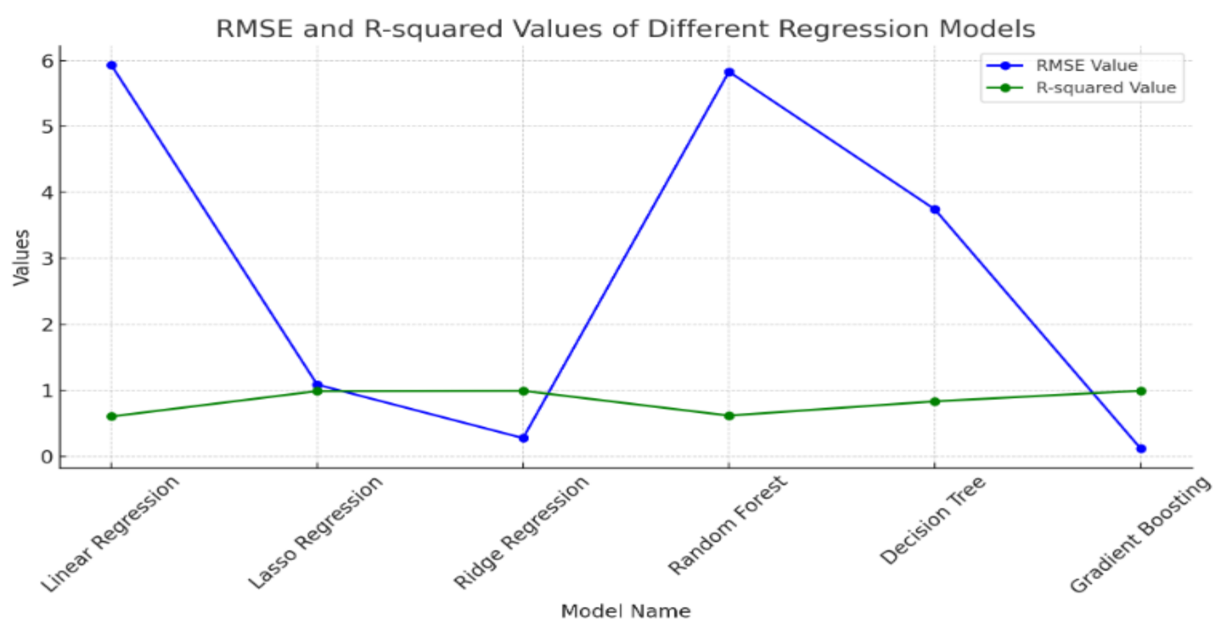

FIGURE NO 8: PLOT FOR MODELS ACCURACY

This experiment demonstrates how machine learning and data analytics may improve the efficiency and dependability of food delivery services. The predicted delivery time feature improves operational efficiency and customer experience greatly, as evidenced by the RMSE and R-squared values of the models utilized. Continuous innovation and flexibility to market developments and customer behavior are required to keep this innovative feature effective.

7. CONCLUSION AND FUTURE SCOPE

Our project has led to the effective development of a predicted delivery time function for a food delivery app, in response to rising consumer demand for prompt and efficient service. We used a comprehensive approach, employing a variety of machine learning models with varied degrees of accuracy and performance. Linear Regression, Ridge Regression, Lasso Regression, Decision Tree, Random Forest, and Gradient Boosting were among the models investigated. From the figure 8 Gradient Boosting was shown to be the most accurate model, with an RMSE of 0.1146 and an R-squared of 0.9925, demonstrating its ability to handle the intricacies of prediction tasks. The use of these predictive models has the potential to greatly improve operational efficiency and customer happiness, providing firms with a considerable competitive advantage in the fast-paced food delivery market.

Moving towards the future, the project intends to improve these models by including additional real-time data and broadening the existing feature set to incorporate other variables that may affect delivery timings, such as driver experience, restaurant efficiency, and customer feedback. Real-time analytics will be critical in responding to urgent concerns such as unexpected weather changes or traffic congestion. To improve model accuracy over time, continuous learning approaches will be used. Furthermore, experimenting with sophisticated machine learning approaches such as deep learning may yield new insights and improve prediction skills. Future efforts will also address the difficulties encountered, such as fluctuations in data quality and the unpredictability of external factors, to assure the dependability of the anticipated delivery time feature and establish it as a standard in the food delivery sector.

8. REFERENCES

- Yaguang Li, Kun Fu, Zheng Wang, Cyrus Shahabi, Jieping Ye, Yan Liu. Multi-task Representation Learning for Travel Time Estimation. In Yike Guo, Faisal Farooq, editors, Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018. pages 1695-1704, ACM, 2018. [doi]
- Chengliang Gao, Fan Zhang, Guanqun Wu, Qiwan Hu, Qiang Ru, Jinghua Hao, Renqing He, Zhizhao Sun. A Deep Learning Method for Route and Time Prediction in Food Delivery Service. In Feida Zhu 0002, Beng Chin Ooi, Chunyan Miao, editors, KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021. pages 2879-2889, ACM, 2021. [doi]

- Z. Chen, H. T. Shen and X. Zhou, "Discovering popular routes from trajectories," 2011 IEEE 27th International Conference on Data Engineering, Hannover, Germany, 2011, pp. 900-911, doi: 10.1109/ICDE.2011.5767890.
- Alajali W, Zhou W, Wen S, et al (2018) Intersection traffic prediction usingdecision tree models. Symmetry 2018, Vol 10, Page 386 10:386.
- Richard Barnes, Senaka Buthpitiya, James Cook, Alex Fabrikant, Andrew Tomkins, Fangzhou Xu. BusTr: Predicting Bus Travel Times from Real-Time Traffic. In Rajesh Gupta 0001, Yan Liu 0002, Jiliang Tang, B. Aditya Prakash, editors, KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020. pages 3243-3251, ACM, 2020. [doi]
- Pourrahmani, Elham & Jaller, Miguel & Fitch-Polse, Dillon. (2023). Modeling the online food delivery pricing and waiting time: Evidence from Davis, Sacramento, and San Francisco. Transportation Research Interdisciplinary Perspectives. 21. 100891. 10.1016/j.trip.2023.100891.
- H. ŞAHİN and İ. Duygu, "Application of Random Forest Algorithm for the Prediction of Online Food DeliveryService Delay," Turkish Journal of Forecasting, vol. 5, no. 1, pp. 1-11. https://doi.org/10.34110/forecasting.842180

  Zhang Y, Haghani A. A gradient boosting method to improvetravel time prediction. Transp Res Part C: Emerg Technol.2015;58:308–24
- Xiaodi Zhu, Rabaa'i, A. A., Jayaraman, J. D., Singh, P., & Mishra, A. (2022). Impact of Social Media on Digital AssetMarkets. 49th Northeast Business & Economics Conference, 138–141.
- Chu H, Zhang W, Bai P, Chen Y. Data-driven optimization for last-mile delivery. Complex & Intelligent Systems. 2023;9(3):2271-2284. https://www.proquest.com/scholarly-journals/data-driven-optimization-last-mile-delivery/docview/2825544853/se-2. doi: https://doi.org/10.1007/s40747-021-00293-1.

  Jie Zheng, Ling Wang, Jingfang Chen, Zi-Xiao Pan, Donghui Li, Yile Liang, Xuetao Ding: A Predictive-Reactive Optimization Framework With Feedback-Based Knowledge Distillation for On-Demand Food Delivery. IEEE Trans. Intell. Transp. Syst. 24(12): 14841-14857 (2023)
- Wang W, Jiang L. Two-stage solution for meal delivery routing optimization on time-sensitive customer satisfaction. J Adv Transport. 2022;2022. https://www.proquest.com/scholarly-journals/two-stage-solution-meal-delivery-routing/docview/2690835192/se-2. doi: https://doi.org/10.1155/2022/9711074.
- Pereira, Veridiana & Saut, Ana & Pieralini, Asta & Albarracin, Orlando. (2023). Online Food-delivery Ecosystem: managing delivery time.
- Liu, Sheng and He, Long and Shen, Zuo-Jun Max, On-Time Last Mile Delivery: Order Assignment with Travel Time Predictors (May 15, 2018). Forthcoming in Management Science, Available at SSRN: https://ssrn.com/abstract=3179994 or http://dx.doi.org/10.2139/ssrn.3179994
- Wu, Chun-Hsin & Ho, Jan-Ming & Lee, D.. (2005). Travel-Time Prediction With Support Vector Regression. Intelligent Transportation Systems, IEEE Transactions on. 5. 276 - 281. 10.1109/TITS.2004.837813.