Food Delivery Time analysis and prediction

Executive Summary

Accurate time estimation is crucial to ensure customer satisfaction and operational efficiency in the growing food delivery industry. This project focuses on analyzing the factors affecting food delivery time and evaluating the performance of machine learning models in estimating delivery time. For this purpose, this project has compiled a detailed dataset from a food delivery company on the Kaggle, including delivery addresses, order times, delivery times, weather conditions, traffic intensity, and delivery person profile information. The project has evaluated the effectiveness and efficiency of various machine learning models such as LightGBM, Random Forest, XGBoost, and k-nearest nearest neighbor (KNN) using metrics such as MAE and RMSE. The results demonstrate that the LGBMRegressor model outperforms other models in accurately predicting delivery times with RMSE and MAE values after optimization of 3.86 and 3.08, respectively. In addition, a thorough analysis of the impact of each feature through a linear regression model has revealed that most of the factors have an impact on delivery times. This project provides insights into leveraging machine learning techniques to optimize food delivery operations and improve customer satisfaction. These findings can assist food delivery platforms in implementing effective time estimation models and highlighting factors for prediction.

Introduction

Food delivery services have expanded significantly in recent years with the advent of online platforms that connect customers, food delivery drivers, and restaurants. One of the challenges that customers face is the uncertainty surrounding delivery times, which has led researchers to develop predictive models that can estimate delivery times, hence food delivery time prediction models will play an important role in this industry. In addition, identifying the factors that will impact delivery times is also essential in building a model. Delivery times directly affect customer satisfaction and their overall experience. The model helps optimize operations, streamline logistics, and meet customer expectations by providing reliable estimates of when food will be delivered.

Objectives

A. Project Objectives

The primary objective of the this project is using data analysis and statistical techniques to build a simple model to identify the factors that affect delivery time, and then propose some solutions based on each factor to improve the time. In addition, we will also build a powerful predictive model to be able to estimate the required delivery time based on the actual factors at that time. These factors can include traffic conditions, time of day, day of week, weather conditions, and even the current workload of delivery partners.

Furthermore, the model can be continuously improved and updated as new data becomes available. By regularly updating the model with the latest delivery data, we can perform some techniques such as Retrain model to ensure that the model remains relevant and reflects the changing conditions and trends in the delivery process.

It is important to note that real-time delivery time prediction requires a dynamic and responsive system. As new orders come in and delivery partners are assigned, the model needs to recalculate and adjust the estimated delivery time based on current circumstances. This ensures that customers receive up-to-date and reliable information about when they can expect their food to arrive.

B. Key Deliverables

The project will deliver the following key deliverables:

- **Data analysis**: This project will provide a detailed analysis of the data collected, including trends, patterns, and correlations identified.
- **Actionable insights**: Our analysis will uncover actionable insights and recommendations to guide strategic decision-making and solve the business problems.
- **Data visualization**: Visualization of key findings will be provided to enhance understanding and facilitate communication of complex data insights.

- **Machine learning model**: The project will provide a machine learning model that predicts the time it takes to deliver a product based on factors that affect the delivery process.

Methodology

A. Data Collection

Data was collected from **Kaggle,** ensuring completeness and accuracy. Contains 20 fields in total and 45593 rows, related to food delivery processes. Variables like the delivery person's identity, age, rating scores, coordinates of the restaurant and delivery point, order and delivery times, weather conditions, traffic intensity, condition of the delivery vehicle, and delivery type are analyzed to understand and predict delivery times. These data are collected to thoroughly examine factors influencing food delivery times and predict them using Machine Learning models.

B. Data Analysis Techniques

Our project employed advanced data analysis techniques, including **Descriptive Analysis, Factor Analysis, Regression Analysis,…** to uncover patterns, trends, and correlations within the data. These techniques were chosen based on the nature of the data and the objectives of the analysis. By leveraging these sophisticated methods, we aimed to extract valuable insights and actionable recommendations to drive informed decision-making and strategic planning, also build a robust model for prediction.

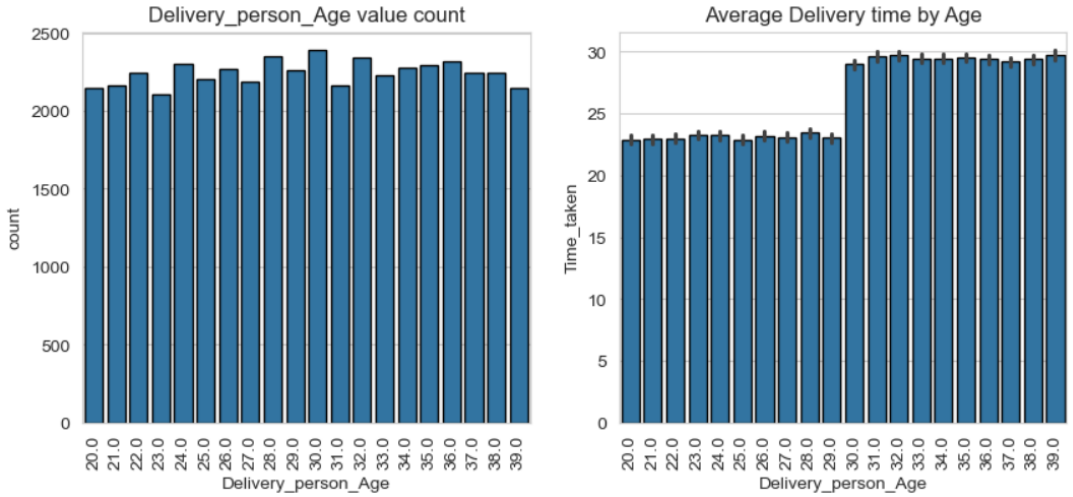
Key Findings

A. Patterns and Insights

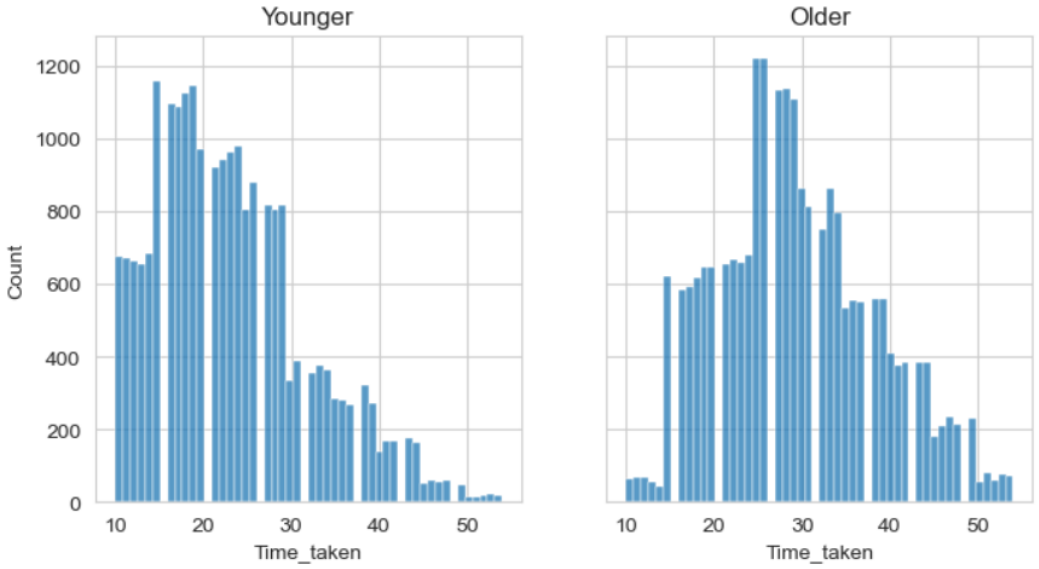
The analysis revealed several key patterns and actionable insights within the data, including **Delivery_person_Age, Delivery_person_Ratings, Distance, Multiple_deliveries and Road_traffic_density**. Although other factors also have some impact, we do not see much difference that the factors bring. However, the following factors still provide us with many valuable insights.

1. ***Delivery_person_Age***

The age of delivery personnel spans a range from 20 to 39 years, with no significant variation in the distribution of individuals across this range. Each age group has a relatively uniform representation, with around 2000 delivery personnel per age. However, despite this even distribution, age demonstrates a clear impact on delivery times. Delivery personnel younger than 30 consistently deliver orders faster than their older counterparts. Specifically, the mean delivery time for those under 30 is approximately 23 minutes, while for those aged 30 or older, the mean delivery time increases to about 29 minutes.

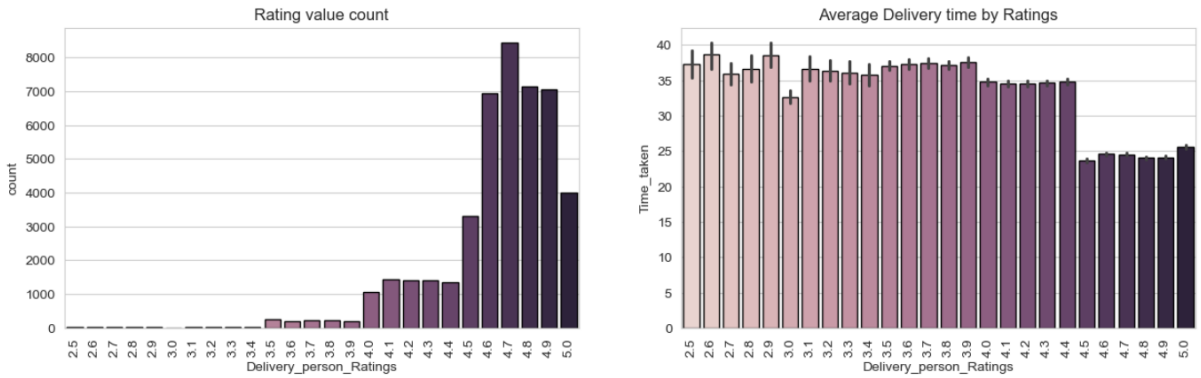


This disparity suggests that younger delivery personnel may possess certain advantages, such as higher physical stamina, quicker navigation skills, or greater adaptability to high-pressure environments. Furthermore, the data reveals that delivery times among the younger group exhibit greater variability, with more outliers indicating occasional instances of exceptionally fast or slow deliveries. In contrast, the older group's delivery times tend to cluster around the higher mean, reflecting more consistent but slower performance.



Interestingly, it appears that older delivery personnel may take on a heavier workload, with a higher likelihood of handling multiple deliveries in a single trip. This added responsibility could contribute to their longer delivery times. Beyond this observation, the overall statistical patterns remain similar across age groups, indicating that while age plays a significant role in delivery efficiency, other factors—such as traffic, distance, or ratings—also interact to influence delivery outcomes.

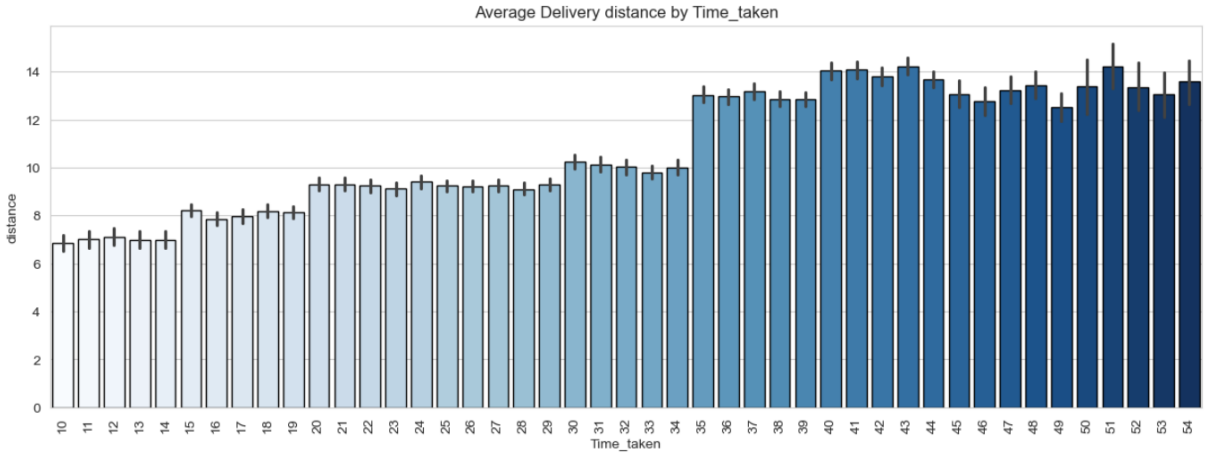
2. ***Delivery_person_Ratings***



Delivery personnel ratings show a noteworthy correlation with delivery efficiency. The majority of delivery personnel have ratings ranging from 3.5 to 4.8, with a smaller subset achieving ratings above 4.8. Higher ratings not only reflect customer satisfaction but also indicate better delivery performance, as personnel with ratings above 4.5 consistently deliver orders faster compared to their lower-rated counterparts. This observation suggests that higher-rated delivery personnel might

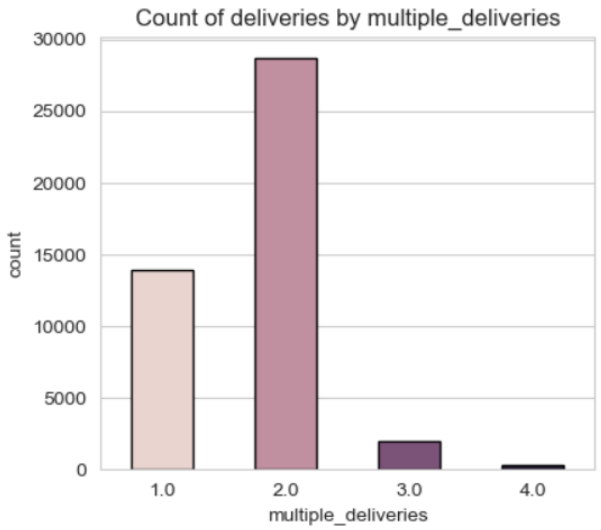
possess superior time management skills, route familiarity, or overall work efficiency. Moreover, it highlights that customer ratings serve as a valuable metric for assessing operational performance and could be used to identify and incentivize high-performing delivery personnel. However, it is also important to note that while high ratings are associated with faster deliveries, other contextual factors such as order volume or distance could influence delivery times and ratings.

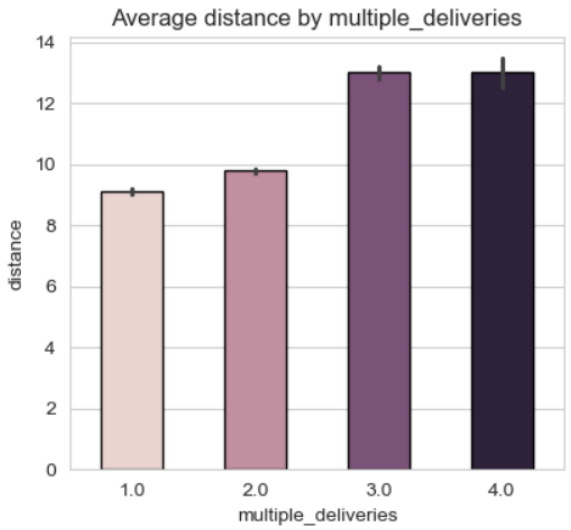
3. ***Distance***



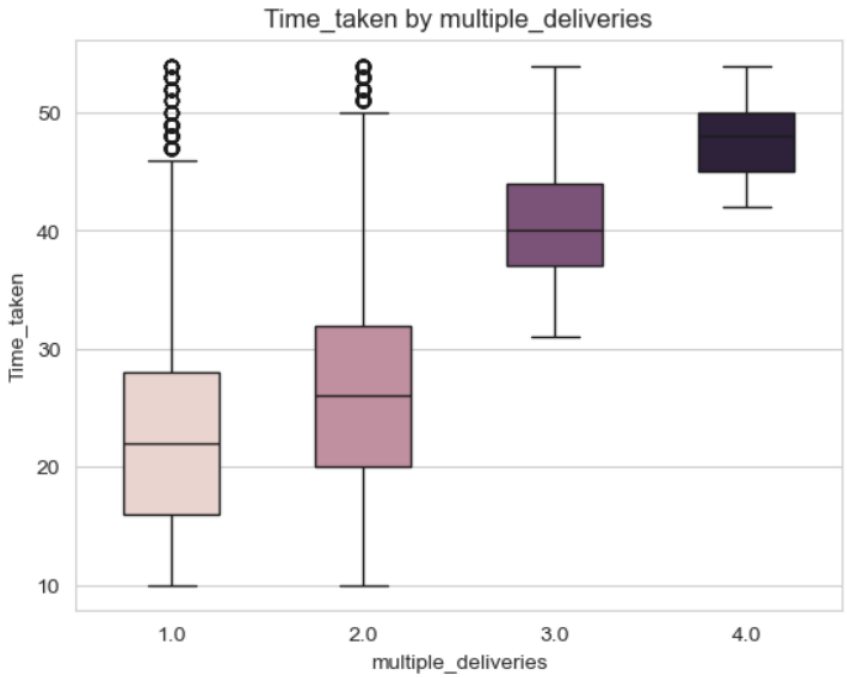
The distance between the restaurant and the delivery destination emerges as a primary determinant of delivery time. Short-distance deliveries, typically under 5 kilometers, are completed in an average time of 18 to 20 minutes, whereas deliveries for distances exceeding 10 kilometers can take upwards of 40 minutes. This finding underscores the inherent challenges of long-distance deliveries, such as increased travel time, traffic variability, and possible delays in locating destinations. Despite

these challenges, the company appears to be optimized for shorter routes, as evidenced by the relatively faster times for deliveries within close proximity. This insight suggests that the company could benefit from exploring strategies to streamline long-distance delivery processes, such as improved route planning or dynamic resource allocation to reduce delays for longer trips.



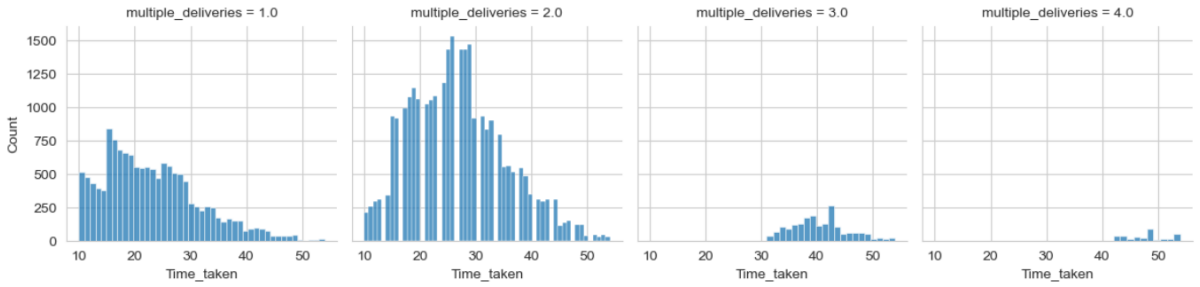


4. ***Multiple_deliveries***

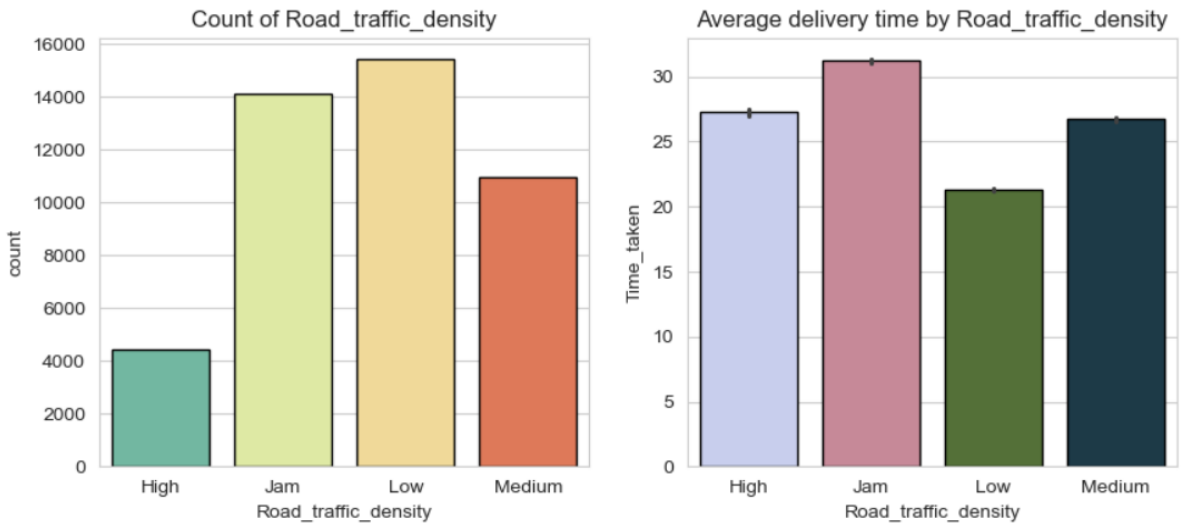


Handling multiple deliveries in a single trip significantly affects the delivery time. Single deliveries, where the delivery partner focuses on just one order, have an average time of around 25 minutes, making them the fastest. However, as the number of deliveries

per trip increases, so does the average delivery time. For example, delivering two or more orders simultaneously leads to noticeably longer times, and this delay becomes even more pronounced when managing more than three orders at once. This pattern likely arises due to the added complexity of navigating between multiple drop-off locations and managing the expectations of several customers simultaneously. While batching multiple deliveries may improve overall efficiency by reducing the number of trips, it also introduces logistical challenges that could negatively impact customer satisfaction for individual orders. To address this, the company might consider strategies such as limiting the number of orders per trip during peak hours or using advanced algorithms to optimize delivery sequences.

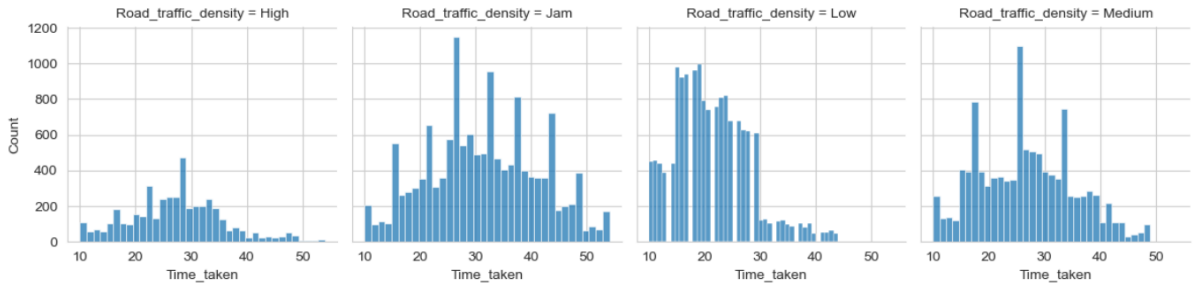


5. ***Road_traffic_density***



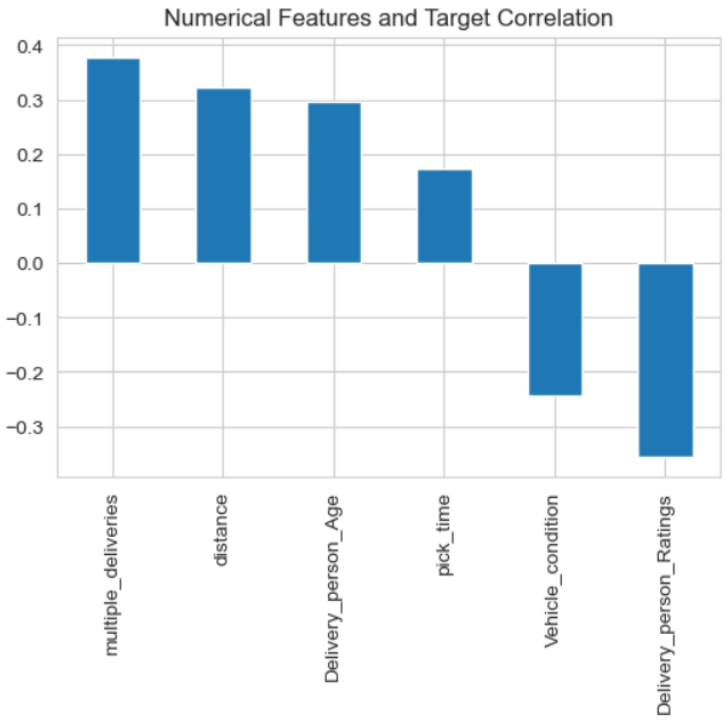
The level of traffic density along delivery routes plays a critical role in determining delivery times. Under low-traffic conditions, delivery times are shortest, averaging approximately 20 minutes. However, as traffic density increases to medium levels, average delivery times rise to around 28 minutes. During periods of high traffic congestion, delivery times become significantly longer, with averages exceeding 35 to

40 minutes. This trend highlights the adverse impact of traffic on delivery efficiency and customer experience, particularly in urban areas prone to congestion during peak hours. The findings suggest that the company could focus on traffic-aware route planning to minimize delays. For example, leveraging real-time traffic data or utilizing alternative delivery methods such as bicycles or motorbikes could prove effective during high-traffic periods. Additionally, incentives for pre-scheduled deliveries during off-peak hours might help mitigate the challenges posed by road congestion.



B. Correlation and Model

Furthermore, correlations between numerical factors and target were identified, leading to insights that **multiple_deliveries and Delivery_person_Ratings** have strong correlation with Delivery time.



In addition, when considering the relationship between the categorical variables and the target variable, we can also see that only the **Type_of_Order** variable has no difference in mean value, or in other words, has no impact on delivery time.

In this project, four different Machine Learning models were employed to predict the food delivery times. Each model has distinct characteristics and offers solutions for

Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken
LGBMRegressor	0.82	0.82	3.93	0.46
XGBRegressor	0.82	0.82	3.96	0.27
RandomForestRegressor	0.81	0.81	4.04	12.88
KNeighborsRegressor	0.75	0.76	4.60	0.56

different types of data structures and complexities. These models include **KNN, Random Forest, XGBoost and LightGBM.**

Results of training between 4 models

After testing with 4 popular models, the results show that the LGBMRegressor model gives the lowest RMSE with default parameters. However, to improve this index, we will perform Hyperparameter tuning to optimize the model.

With the final model, we test once on the test set and evaluate the model's ability to respond to an unseen dataset.

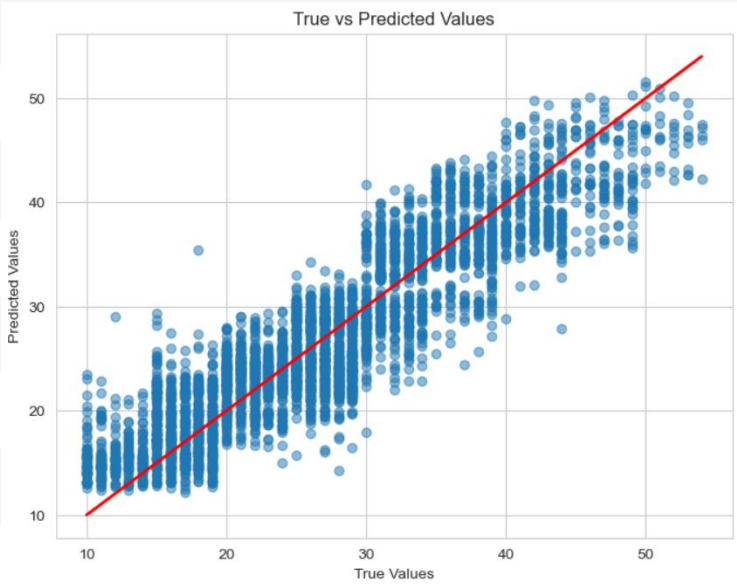
	MSE	RMSE	MAE	Explained Variance	Max Error
Score	15.41	3.93	3.16	0.82	17.44

Model best result

Model prediction statistics

Looking at the data statistics, we see that the results between the actual value and the

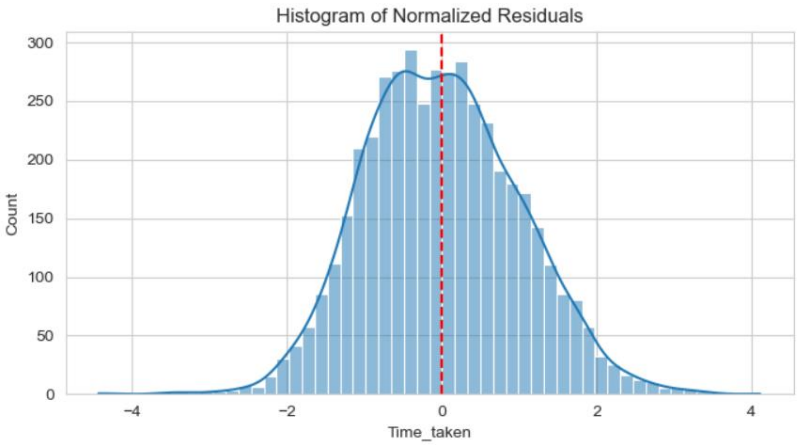
	Real Value	Model Prediction	Difference	Difference %
count	4490.00	4490.00	4490.00	4490.00
mean				13.61
std				12.72
min				0.00
25%				5.26
50%				11.11
75%				18.41
max				141.67



predicted value are quite similar to each other. However, most of the models do quite well when most of the differences are less than 3 minutes. However, there is still a difference of up to 17 minutes compared to the actual value.

Model prediction values vs Real values

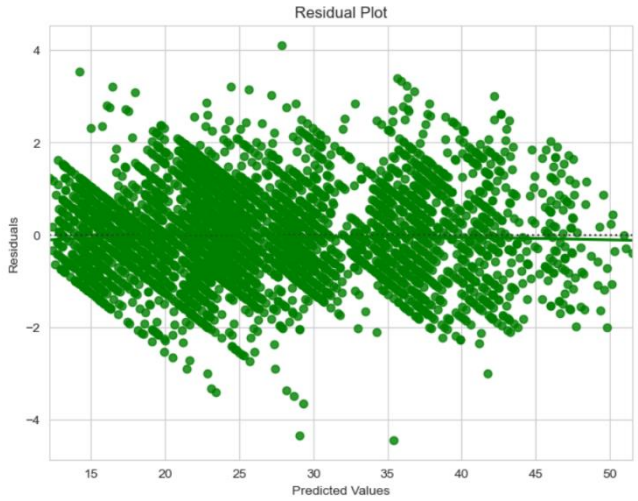
The above graph shows a comparison between the predicted values and the actual values. The results show that the values are quite close to the diagonal (the diagonal y=x shows the match between the predicted values and the actual values). In addition,



the distribution of the points is quite even and random around the diagonal, which shows that the model has good generalization ability.

Distribution of Normalized Residuals

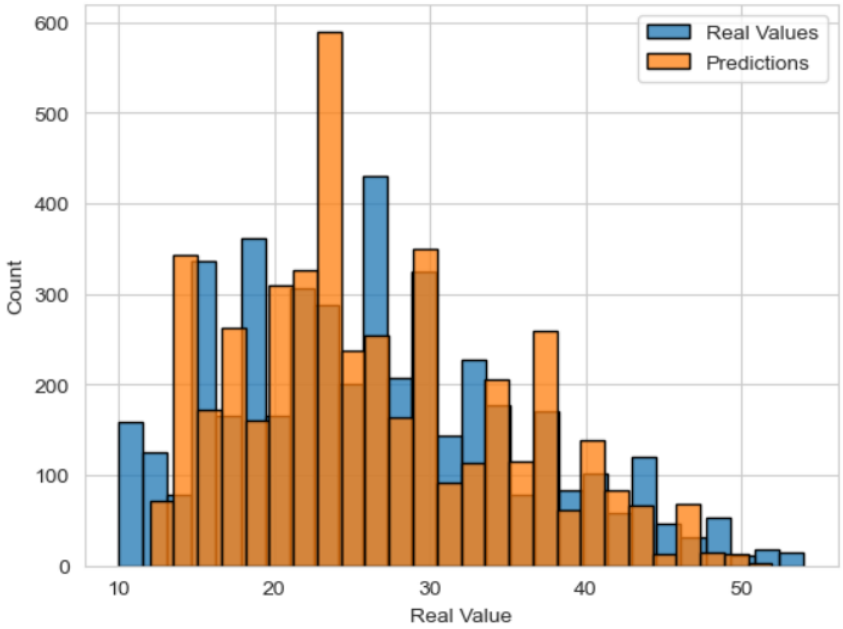
The above graph is used to evaluate the normal distribution of the residuals. It can be seen that the residuals have approximately normal distribution when they have a bell



shape around the 0 axis, thereby showing that the hypothesis of a normally distributed residual model is not violated, ensuring the model is suitable.

Residual plot

The above graph helps to evaluate the validity of the model and identify potential problems. One of the assumptions of regression is homoscedasticity. This means that the dispersion of the residuals does not change with the predicted values. With the above results, the values are around the horizontal line 0 and there is no increasing



or decreasing trend, indicating that the model is performing well. However, the results also show that there are still some outliers that should be removed to achieve better results.

Distribution of Predicted values and Real values

The distribution of predictions and actual values was found to be similar, suggesting that the model's predictions aligned well with the real data.

Conclusion

In summary, this project successfully leveraged various statistical and machine learning techniques to solve a real-world problem in the context of express delivery companies. The goal was to analyze the influencing factors and develop an effective prediction model of delivery time by discovering hidden patterns and relationships in a large dataset.

The project followed a well-defined workflow, including data collection, cleaning and pre-processing, exploratory data analysis, model building, and evaluation. Each stage was approached systematically, using tools such as Python and visualization libraries to facilitate the analysis and modeling process.

As the project came to an end, there were many opportunities for further optimization and refinement. By incorporating additional data and exploring advanced techniques, model performance can be improved, leading to more accurate predictions and actionable insights for Delivery companies.

In summary, this project successfully addresses the problem of delivery time prediction, highlights the value of data science methods, and demonstrates the benefits of a modular architecture for implementation. The insights gained from this project can inform decision making and contribute to improving the operations of today's express delivery companies.