FOOD DELIVERY TIME PREDICTION
DATA ANALYSIS REPORT

I. Introduction

1. Project discription

In the last few years, food delivery has experienced major expansion with the inception of the online platforms that connect customers, delivery drivers and restaurants. One of the difficulties that customers face is the uncertainty surrounding delivery times that prompts researchers to come up with predictive models that will allow estimations of delivery duration, so that the food delivery time prediction model will play a vital role in this industry. Delivery time directly affects customer satisfaction and their overall experience. The model helps optimise operations, streamline logistics, and meet customer expectations by providing reliable estimates of when the food will be delivered.

In this project, I will be performing an supervised machine learning - Regression on a dataset contains the factors that can be affect to food delivery time. I will build a regression model to estimate the time it will take a driver to deliver to the destination.

2. Business task

The main objective of this project is using regression algorithms to predict time taken to delivery based on identify the factors that can affect to food delivery time. This project can enhance customer satisfaction by providing more accurate delivery time estimates, besides, driver allocation can be optimized, reducing idle time and improving resource use.

II. Analysis

1. Data overview

This dataset for analysis was downloaded from Kaggle, 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.

Link dataset:

https://www.kaggle.com/datasets/gauravmalik26/food-delivery-dataset

2. Tools

This project mainly uses Python programming language with some libraries like Pandas, Numpy, Matplotlib, Seaborn, Sklearn for data import, data cleaning, data analysis, visualization and build model.

3. Analysis

See full analysis project here: https://github.com/hoan110102/Food-Delivery-Time-Prediction/blob/main/food_delivery_time.ipynb

4. Result

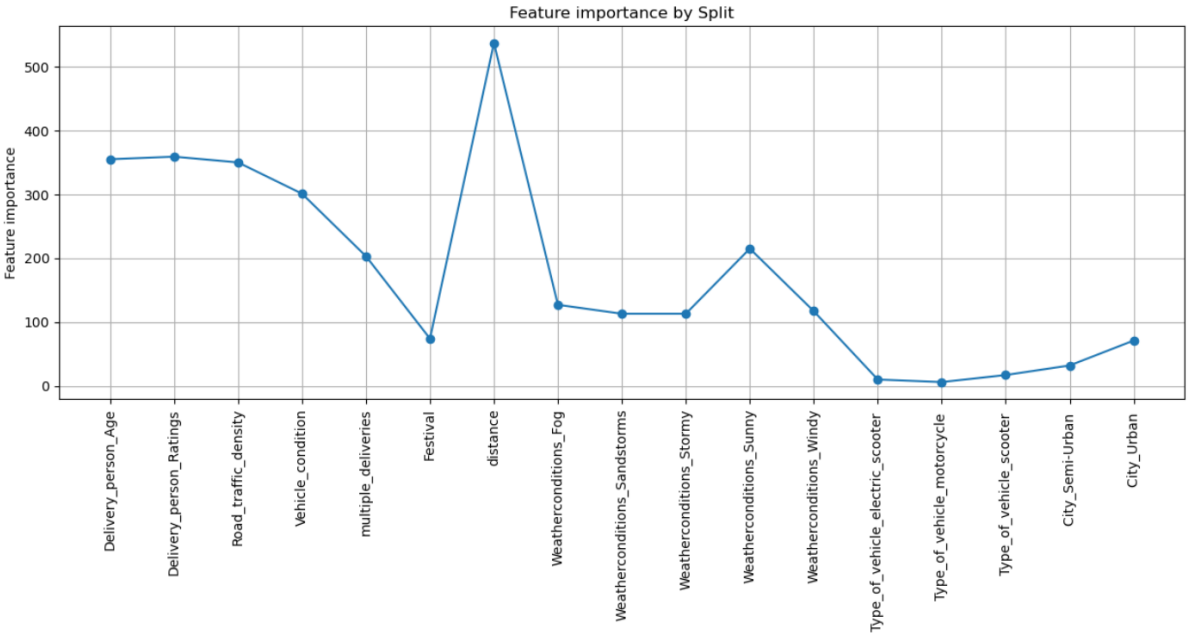
In this project, five different Machine Learning models were employed to predict the food delivery times. Each model has distinct characteristics and offers solutions for different types of data structures and complexities. These models include Linear Regression, Decision Tree, Random Forest, XGBoost, LightGBM.

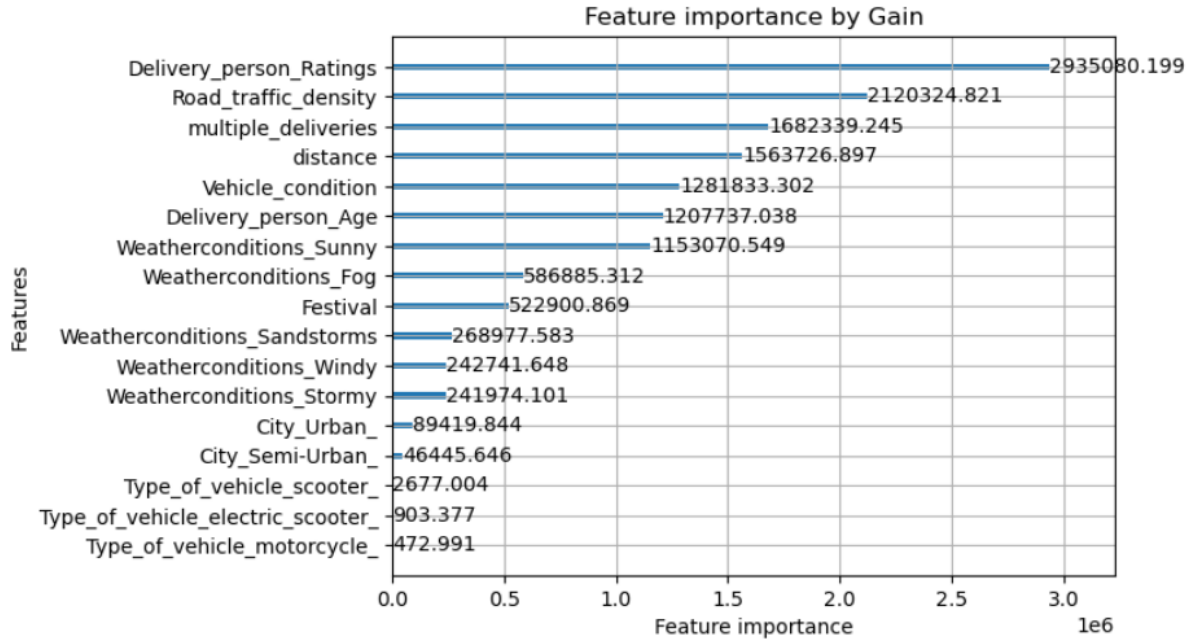
- **Model training result:**

Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken
LGBMRegressor	0.82	0.82	4.01	0.44
XGBRegressor	0.81	0.82	4.04	0.31
RandomForestRegressor	0.80	0.80	4.17	12.22
DecisionTreeRegressor	0.64	0.64	5.62	0.25
LinearRegression	0.52	0.52	6.51	0.10

From the above model training results, we see that the LGBMRegressor model gives the largest R^2 result, we will use this model as the main model.

- **Model evaluation**

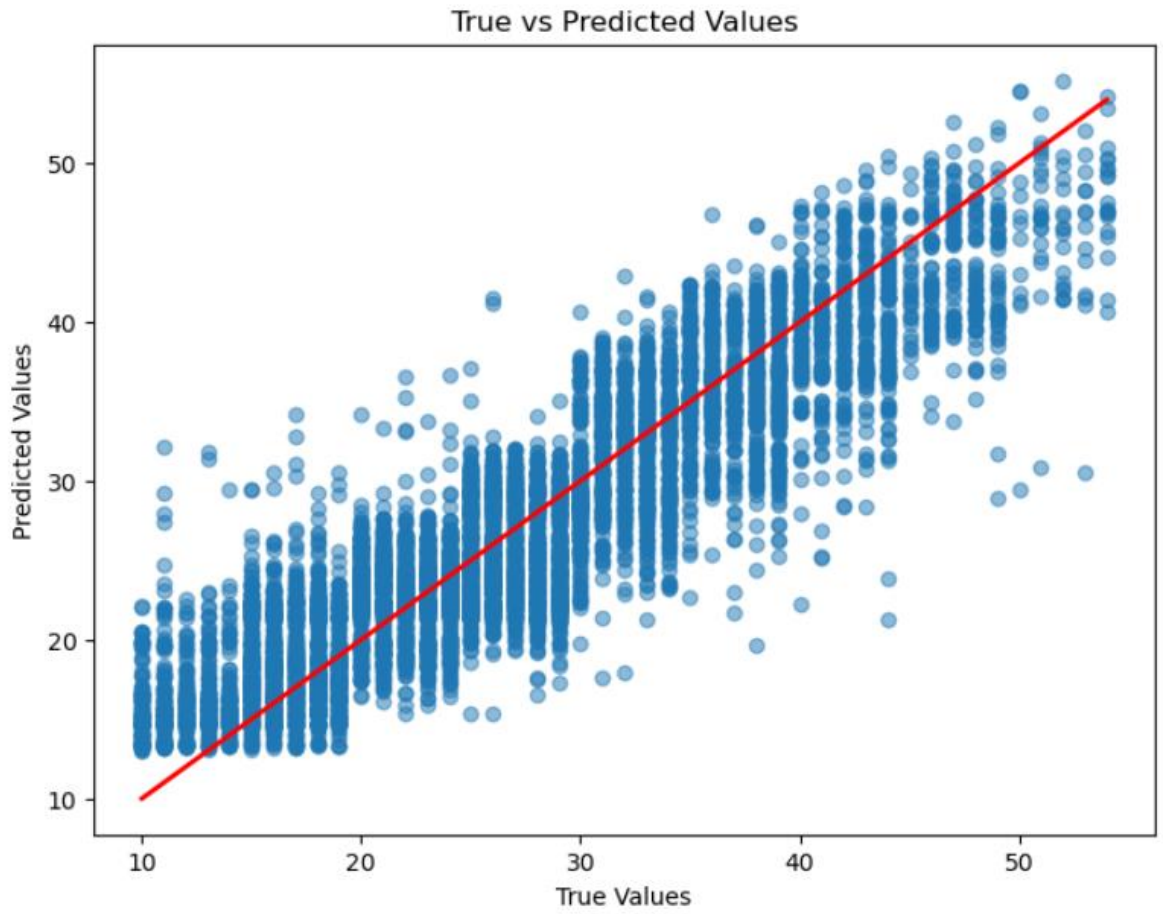




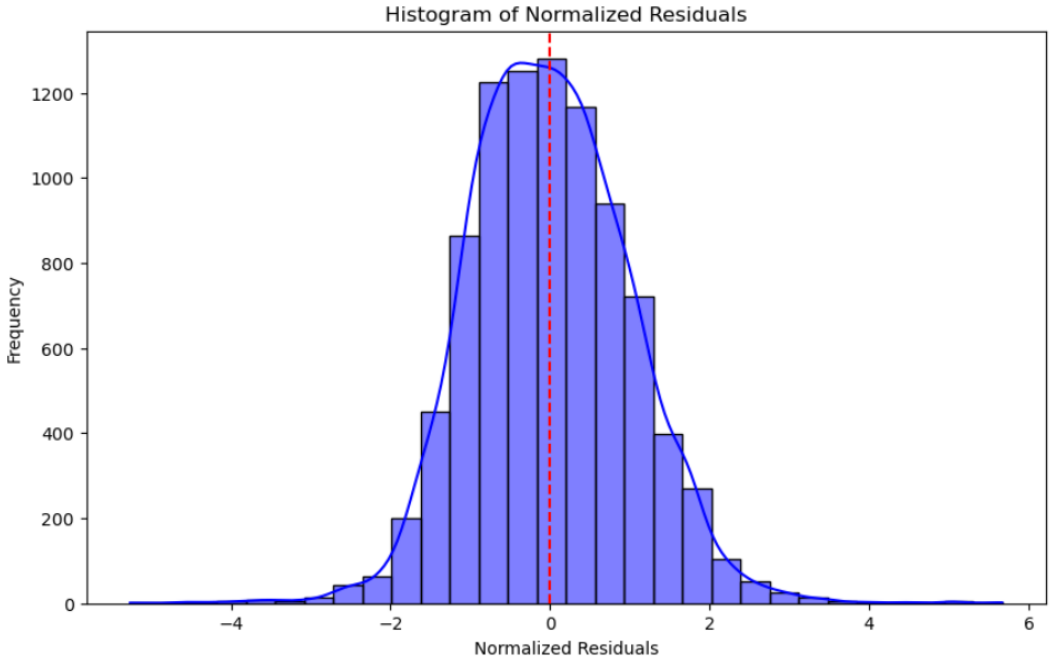
The two graphs above show the results of the importance of each variable in the model in two different ways. With graph 1, it is based on Split, which means counting the number of times a variable is selected to split the data in the trees. The variable that is selected more often tends to be considered more important, because it tends to help classify the data more times. With graph 2, it is based on Gain, which means the amount of improvement that each variable contributes to the model when it is selected to split

the data. Gain measures the importance based on how much better the variable helps the model predict.

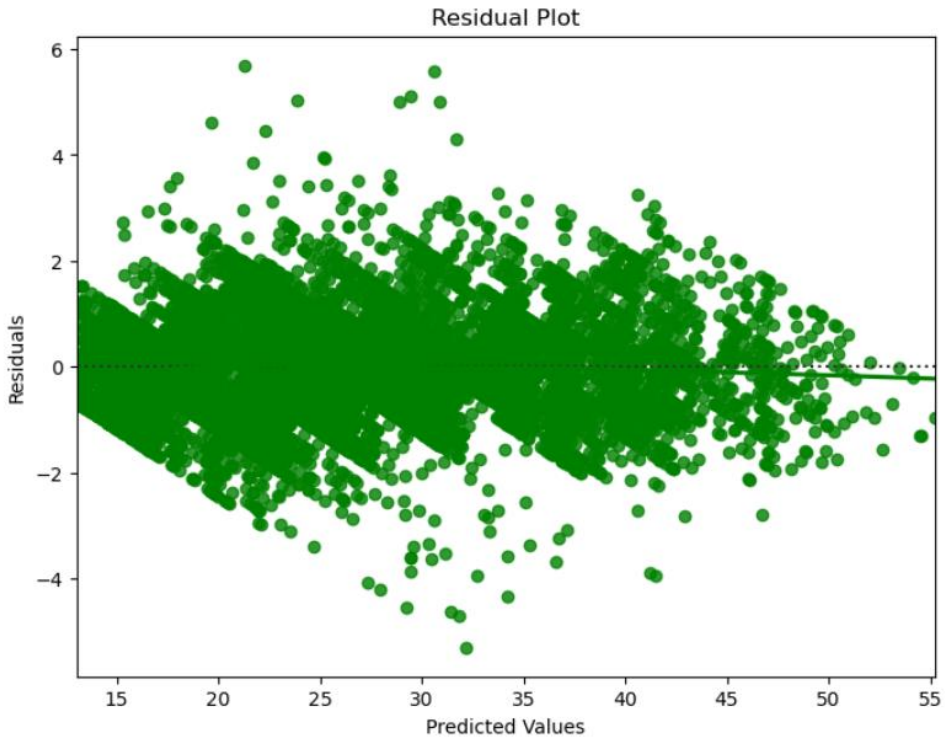
With the above results, although the number of times chosen to divide is the Distance variable, but to reflect the highest contribution of the variables to the model, the Delivery_person_Ratings variable accounts for the most, this is also quite true to reality because the delivery time mostly affects the delivery person's rating, so a person with a high rating means that person delivers quickly (ie low Time_taken) will satisfy customers.



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.



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.



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.