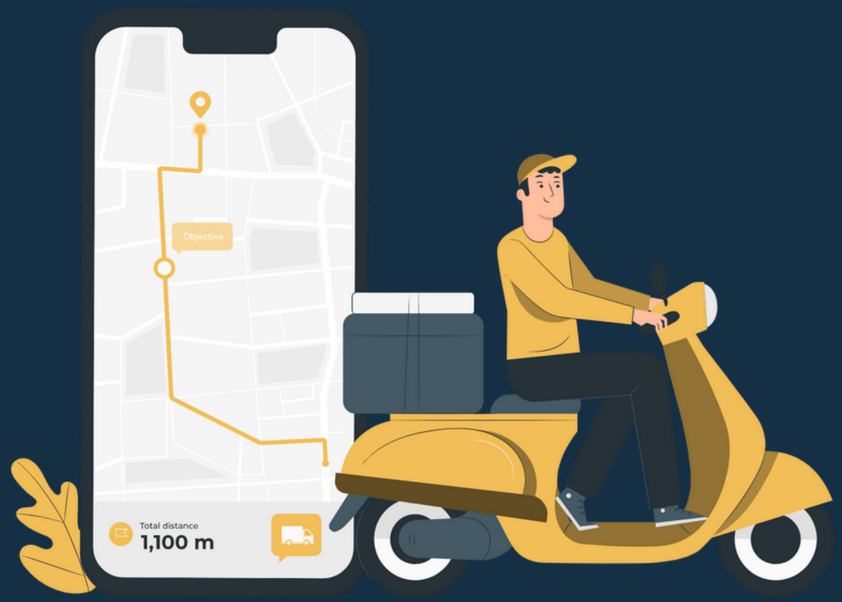


Food Delivery Time Prediction

By: Aldimeola Alfarisy

Background

Delivery service is a form of service that provides services to deliver orders (especially goods) that ordered by customers to a place according to their wishes. Currently, delivery service is one of the services most needed by people in obtaining the goods needed because it saves a lot of time and energy.

Currently there are many types of businesses that have delivery services as one of the services that can be provided to customers. One of them is a business engaged in the food or beverage sector that has a service to deliver food ordered to the place the customer wants.

The timeliness required in delivering food to the customer's place is the main challenge in this service. Accuracy time in delivering the food must be shown to keep transparency with their customers. So, by using historical data on the time it takes to deliver food, the use of machine learning algorithms is one way to predict the accuracy of the time needed to deliver food to location.

Objectives

1. What factors can affect the time in delivering food from the restaurant to the destination location?
2. How much food delivery time prediction accuracy performance?

Data Preparation

General Info

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45593 entries, 0 to 45592
Data columns (total 11 columns):
 #   Column                       Non-Null Count  Dtype  
---  ------                       --------------  -----  
 0   ID                           45593 non-null  object 
 1   Delivery_person_ID           45593 non-null  object 
 2   Delivery_person_Age          45593 non-null  int64  
 3   Delivery_person_Ratings      45593 non-null  float64
 4   Restaurant_latitude          45593 non-null  float64
 5   Restaurant_longitude         45593 non-null  float64
 6   Delivery_location_latitude   45593 non-null  float64
 7   Delivery_location_longitude  45593 non-null  float64
 8   Type_of_order                45593 non-null  object 
 9   Type_of_vehicle              45593 non-null  object 
 10  Time_taken(min)              45593 non-null  int64  
dtypes: float64(5), int64(2), object(4)
memory usage: 3.8+ MB
```

- Dataset consists 45593 rows and 11 columns. Then the dataset also consists of 7 numerical data and 4 categorical data
- The problem faced is a regression problem, namely predicting the time needed to deliver food (Time taken)
- There are no missing value, duplicated value, and odd data in this dataset
- All numerical values contained in the dataset are quite reasonable

Data Preparation

Distance Calculation

```
In [8]: # Set the earth's radius (in kilometers)
        R = 6371

        # Convert degrees to radians

        def deg_to_rad(degrees):
            return degrees * (np.pi/180)

        # Function to calculate the distance between two points using the haversine formula

        def dist_calculate(lat1, lon1, lat2, lon2):
            d_lat = deg_to_rad(lat2-lat1)
            d_lon = deg_to_rad(lon2-lon1)
            a = np.sin(d_lat/2)**2 + np.cos(deg_to_rad(lat1)) * np.cos(deg_to_rad(lat2)) * np.sin(d_lon/2)**2
            c = 2 * np.arctan2(np.sqrt(a), np.sqrt(1-a))
            return R * c

        # Calculate the distance between each pair of points

        data['Distance'] = np.nan

        for i in range(len(data)):
            data.loc[i, 'Distance'] = dist_calculate(data.loc[i, 'Restaurant_latitude'],
                                                     data.loc[i, 'Restaurant_longitude'],
                                                     data.loc[i, 'Delivery_location_latitude'],
                                                     data.loc[i, 'Delivery_location_longitude'])
```

	ID	Delivery_person_ID	Delivery_person_Age	Delivery_person_Rating	Distance	Type_of_order	Type_of_vehicle	Time_taken(min)
0	4607	INDORES13DEL02	37	4.	3.025149	Snack	motorcycle	24
1	B379	BANGRES18DEL02	34	4.	20.183530	Snack	scooter	33
2	5D6D	BANGRES19DEL01	23	4.	1.552758	Drinks	motorcycle	26
3	7A6A	COIMBRES13DEL02	38	4.	7.790401	Buffet	motorcycle	21
4	70A2	CHENRES12DEL01	32	4.	6.210138	Snack	scooter	30

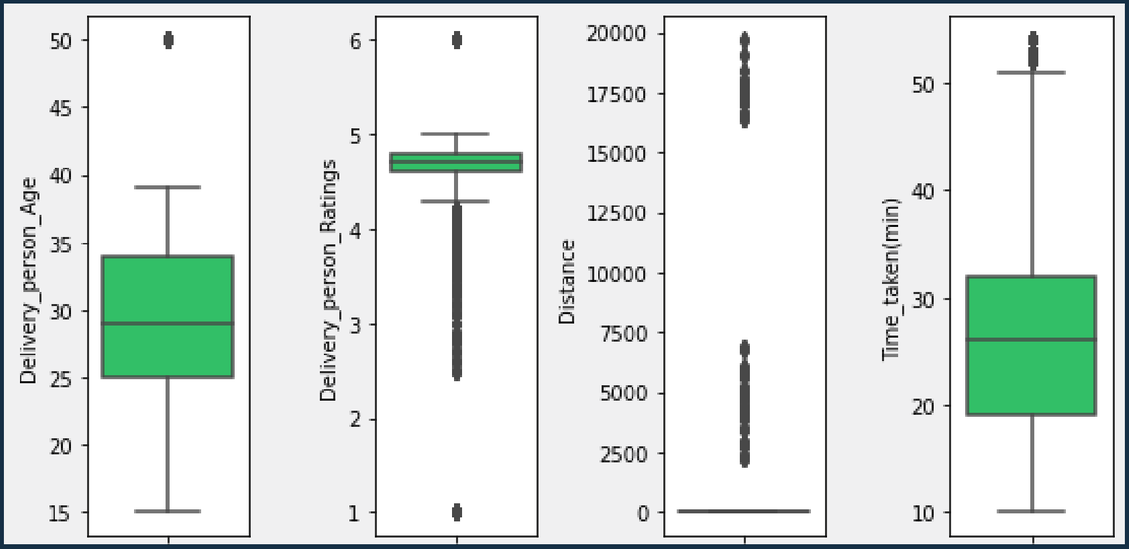
To get the time needed to deliver food, the distance between the restaurant and the delivery location is needed. To get the required distance, **Haversine Formula** can be used to calculate the distance between 2 locations by utilizing longitudes and latitudes.

Haversine Formula

$$
\begin{aligned}
d &= 2r \arcsin\left(\sqrt{\operatorname{hav}(\varphi_2 - \varphi_1) + (1 - \operatorname{hav}(\varphi_1 - \varphi_2) - \operatorname{hav}(\varphi_1 + \varphi_2)) \cdot \operatorname{hav}(\lambda_2 - \lambda_1)}\right) \\
&= 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \left(1 - \sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) - \sin^2\left(\frac{\varphi_2 + \varphi_1}{2}\right)\right) \cdot \sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \\
&= 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos\varphi_1 \cdot \cos\varphi_2 \cdot \sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right).
\end{aligned}
$$

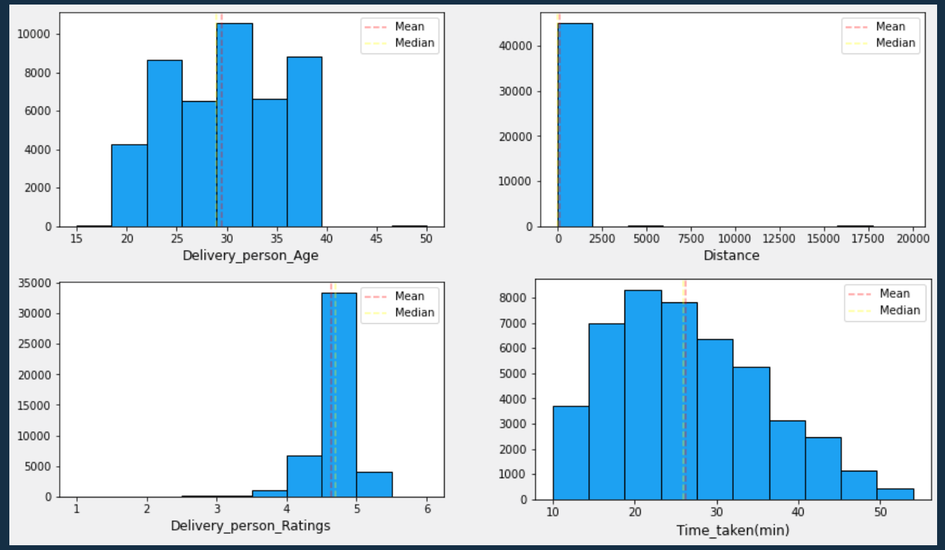
Exploratory Data Analysis

Univariate Analysis



Outliers Check

Extreme outliers at Delivery_person_Ratings column are on lower boundary (< 3.9), meanwhile distance column have extreme outliers in upper boundary (> 31.9)

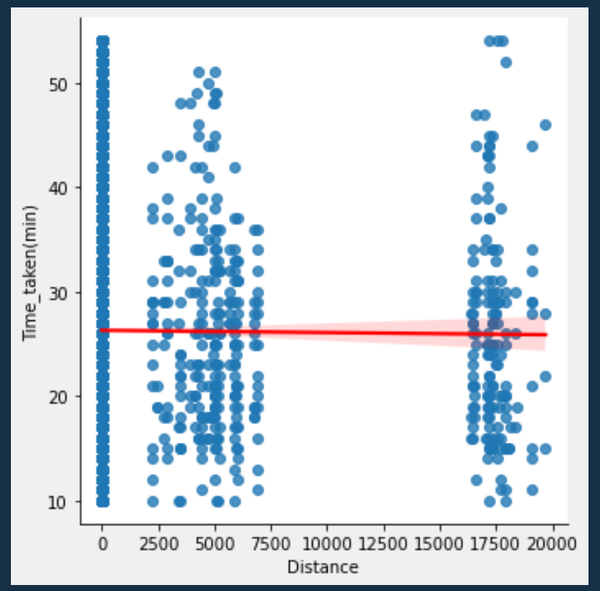


Data Distribution Check

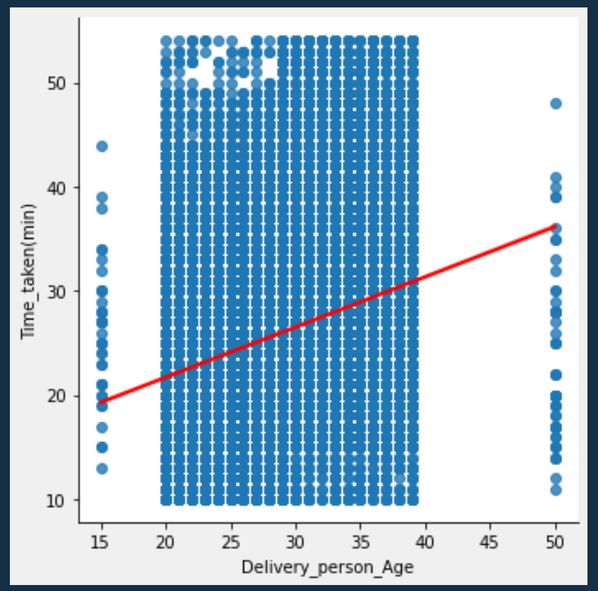
- Delivery_person_Age and Time_taken(min) column have relatively normal data distribution
- Delivery_person_Ratings has negative skew data distribution, meanwhile distance column has positive skew data distribution

Exploratory Data Analysis

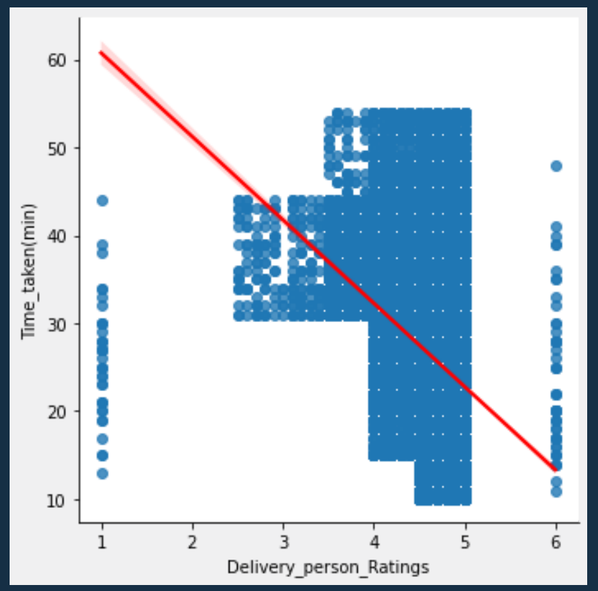
Bivariate Analysis



There is consistent relationship between the time taken and the distance travelled to deliver the food. It looks like majority food delivered within 25-27 minutes regardless of distance



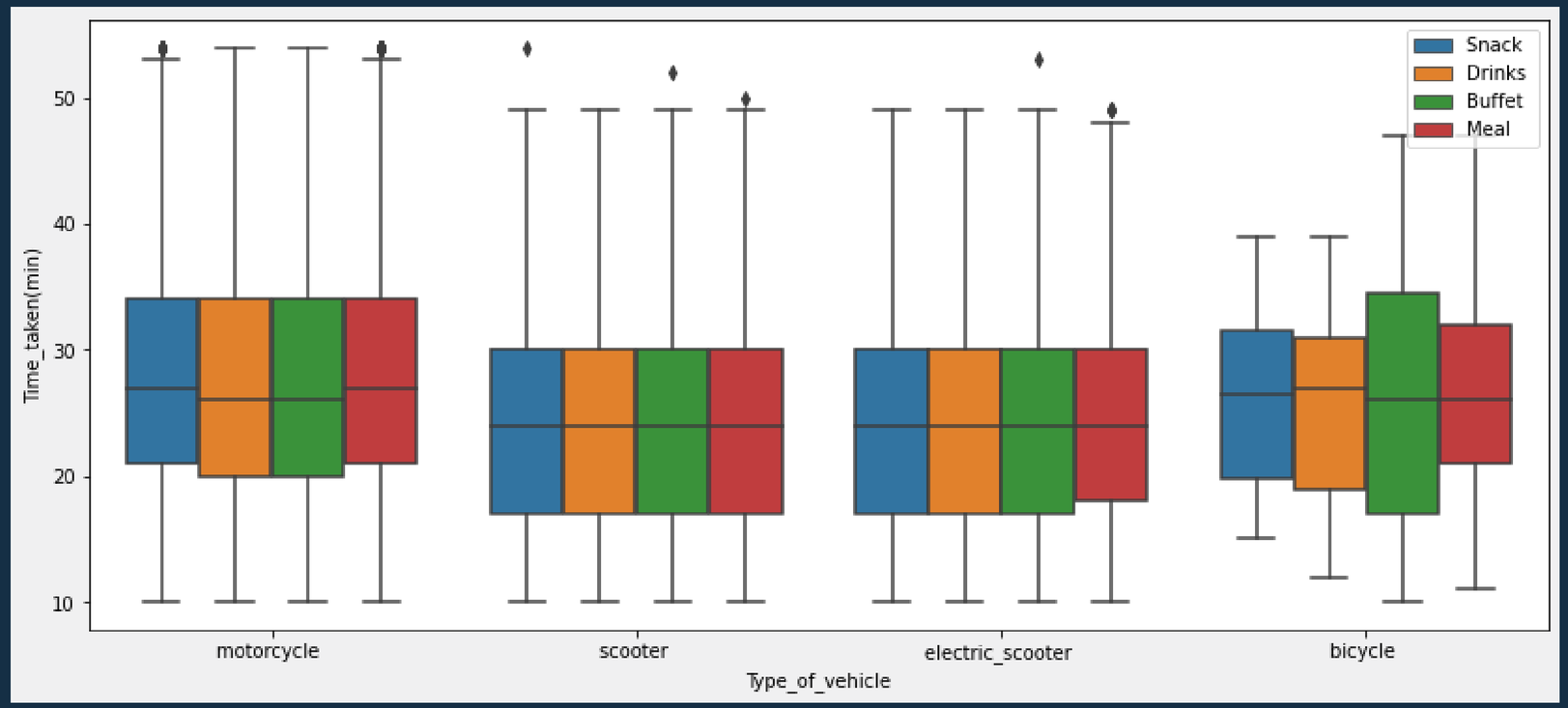
There is a linear relationship between the time taken to deliver the food and the age of the person who delivering the food. It looks like person with the young age able to take less time than person with old age to deliver the food to customers



There is an inverse linear relationship between the time taken to deliver the food and the delivery person ratings. It looks like person with the higher ratings take a less time to deliver the food than person with low ratings

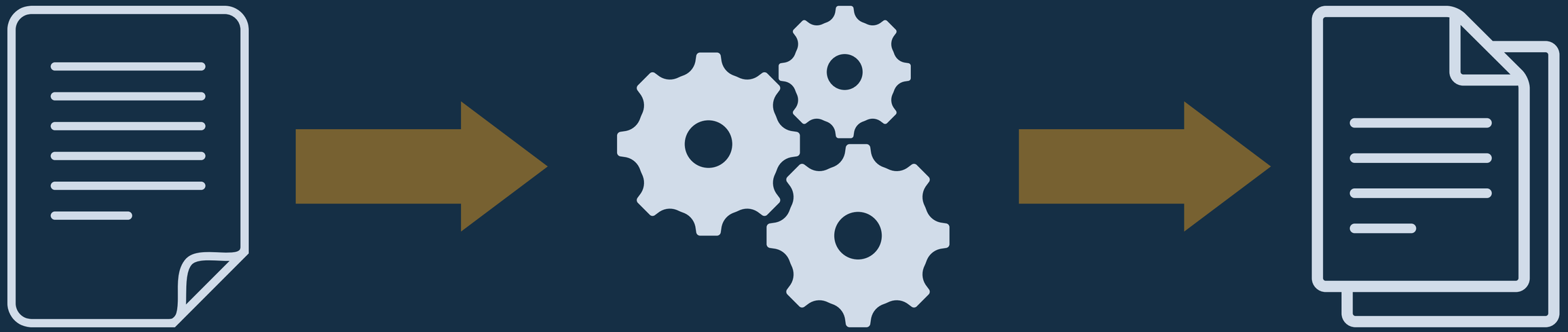
Exploratory Data Analysis

Bivariate Analysis



It looks like there is not much difference between the time taken depending on the vehicle they are driving and the type of food they are delivering

Feature Engineering



Initial Dataset

Feature Engineering

- Unused Feature Drop
- Outliers Handling
- One Hot Encoding
- Train Test Data Split (80:20)

Final Dataset (Train & Test Data)

Modeling and Evaluation

Algorithms Training

	Algorithm	RMSE
0	LinearRegression	7.927093
1	Ridge	7.927084
2	Lasso	8.518059
3	DecisionTreeRegressor	10.291782
4	RandomForestRegressor	7.775025

Based on each **RMSE (Root Squared Mean Error)** result from 5 trained algorithms, best algorithm is random forest regressor because have the smallest error than other algorithms. Performance of the best model will try to be improved with hyperparameter tuning

RMSE (Root Squared Mean Error) Formula

$$RMSE = \text{sqrt}\left(\frac{\Sigma(\text{actual} - \text{prediction})^2}{\text{Number of observations}}\right)$$

Modeling and Evaluation

Hyperparameter Tuning

```
parameters = {'n_estimators': (100, 200, 300, 400, 500),
              'max_depth':(10, 20, 30, 40, 50),
              'bootstrap':('True', 'False')
             }
```

```
{'bootstrap': 'True', 'max_depth': 10, 'n_estimators': 300}
```

```
RMSE:  7.281613193317689
```

Parameters to be tuned

Parameter tune using GridSearch CV

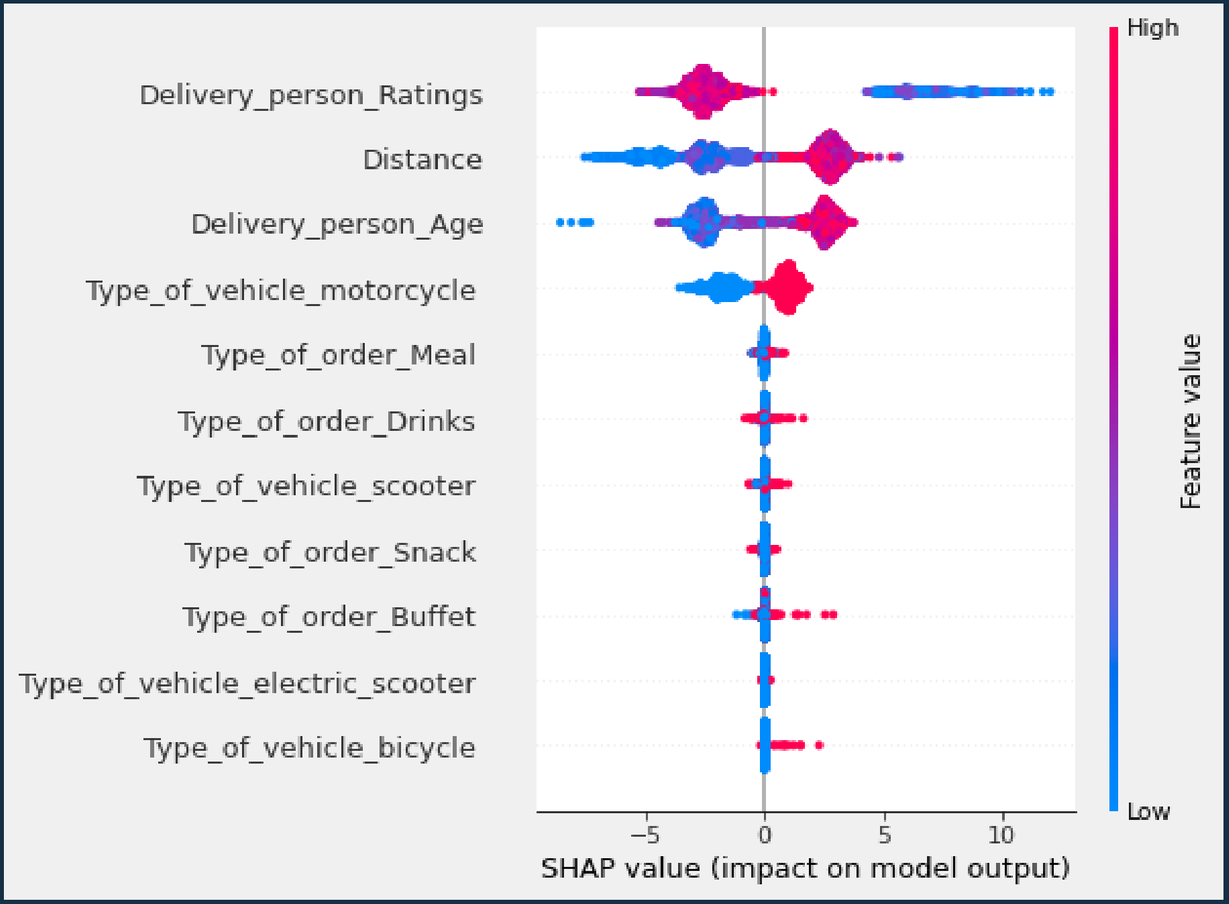
Best parameter

Model train using best parameter

Final model performance

Modeling and Evaluation

Feature Importances



Delivery_person_Ratings variable become the most importance feature in this model, followed by Distance and Delivery_person_age variables. Then, we can see Delivery_person_Ratings variable has a negative contribution when its values are high, and a positive contribution on low values

Prediction on New Data

```
Age of Delivery Person: 30
Rating of Previous Deliveries: 2.3
Total Distance: 16
Order Type: Buffet
Vehicle: electric scooter
Predicted Delivery Time in Minutes:  37
```

Conclusion and Recommendation

Conclusion

1. Rating of person in previous delivers become is the most influential factor on the delivery time of food to the destination location. Person with the higher ratings take a less time to deliver the food than person with low ratings
2. Model has RMSE score 7.28 and that means error between delivery time prediction and delivery time actual is 7.28 minutes

Recommendation

The rating obtained by the deliveryman is a representation of the deliveryman's performance in delivering food to the intended location in terms of delivery time. Of course this is a potential loss of customers if this continues to happen. Delivery time performance needs to be maintained so that the rating obtained is high and customer trust can still be maintained.
The RMSE value of the model can be used as a guarantee of delivery time performance which can be given to the customer so that as much as possible the delivery time is not more than the existing RMSE score (delay in delivery time of not more than 7.28 minutes).

Documentation : **Github Repository**

THANK YOU

+62 878 7375 6695

aldimeolaalfarisy@yahoo.com

aldimeolaa