Hackerearth-amazon-business-research-analyst-hiring-challenge

Build a machine learning model that can calculate the time the delivery person takes to deliver the order.

- Basic exploratory data analysis using pandas, matplotlib, seaborn packages.
- Data pre-processing
    - Missing value indicator
    - Missing value imputation for the columns,
        - delivery_person_age
        - delivery_person_ratings
        - time_orderd
        - weather_conditions
        - road_traffic_density
        - multiple_deliveries
        - festival
        - city
    - Feature Engineering
        - time_order_picked_new
        - diff_order_picked
        - median_diff_order_picked
        - time_orderd_new

        - delivery_person_loc
        - restaurant_delivery_distance
    - Datetime feature engineering

- The final features for the model
    - 0_delivery_person_id
    - 1_delivery_person_age
    - 2_delivery_person_ratings
    - 3_restaurant_latitude
    - 4_restaurant_longitude
    - 5_delivery_location_latitude
    - 6_delivery_location_longitude
    - 7_weather_conditions
    - 8_road_traffic_density
    - 9_vehicle_condition
    - 10_type_of_order
    - 11_type_of_vehicle
    - 12_multiple_deliveries
    - 13_festival
    - 14_city
    - 15_time_taken_(min)
    - 16_delivery_person_age_isnull
    - 17_delivery_person_ratings_isnull
    - 18_time_orderd_isnull
    - 19_weather_conditions_isnull

- 20_road_traffic_density_isnull
- 21_multiple_deliveries_isnull
- 22_festival_isnull
- 23_city_isnull
- 24_time_order_picked_new
- 25_diff_order_picked
- 26_median_diff_order_picked
- 27_time_orderd_new
- 28_delivery_person_loc
- 29_restaurant_delivery_distance
- 30_day
- 31_day_number
- 32_month_number
- 33_year_quarter
- 34_week_of_year
- 35_year
- 36_dayofmonth
- 37_dayofyear
- 38_weekday
- 39_weekend
- 40_month_start
- 41_month_end
- 42_quarter_start
- 43_quarter_end
- 44_year_start

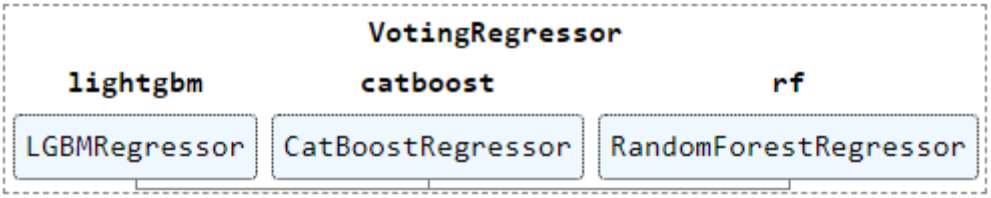
- 45_year_end
- 46_seasons
- 47_timings

- By using pycaret regressor compared more than one regressor model with 5-fold cross-validation and evaluated by the r2 score.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
lightgbm	Light Gradient Boosting Machine	3.1645	15.8566	3.9820	0.8199	0.1633	0.1378	2.7180
catboost	CatBoost Regressor	3.1909	16.0788	4.0098	0.8174	0.1642	0.1389	11.8940
rf	Random Forest Regressor	3.1938	16.3634	4.0451	0.8141	0.1664	0.1389	53.6480
et	Extra Trees Regressor	3.2054	16.8154	4.1006	0.8090	0.1675	0.1386	237.5180
xgboost	Extreme Gradient Boosting	3.2693	16.8733	4.1076	0.8084	0.1686	0.1425	3.9860
gbr	Gradient Boosting Regressor	3.6651	21.1780	4.6018	0.7594	0.1902	0.1626	46.9760
dt	Decision Tree Regressor	4.1180	29.6372	5.4440	0.6634	0.2231	0.1766	3.1760
br	Bayesian Ridge	4.7238	34.9382	5.9108	0.6031	0.2409	0.2083	17.0120
lar	Least Angle Regression	4.7248	34.9697	5.9134	0.6028	0.2406	0.2086	2.0340
omp	Orthogonal Matching Pursuit	4.7482	35.3319	5.9440	0.5986	0.2425	0.2093	1.0840
ridge	Ridge Regression	4.8126	36.2666	6.0221	0.5880	0.2461	0.2120	0.3540

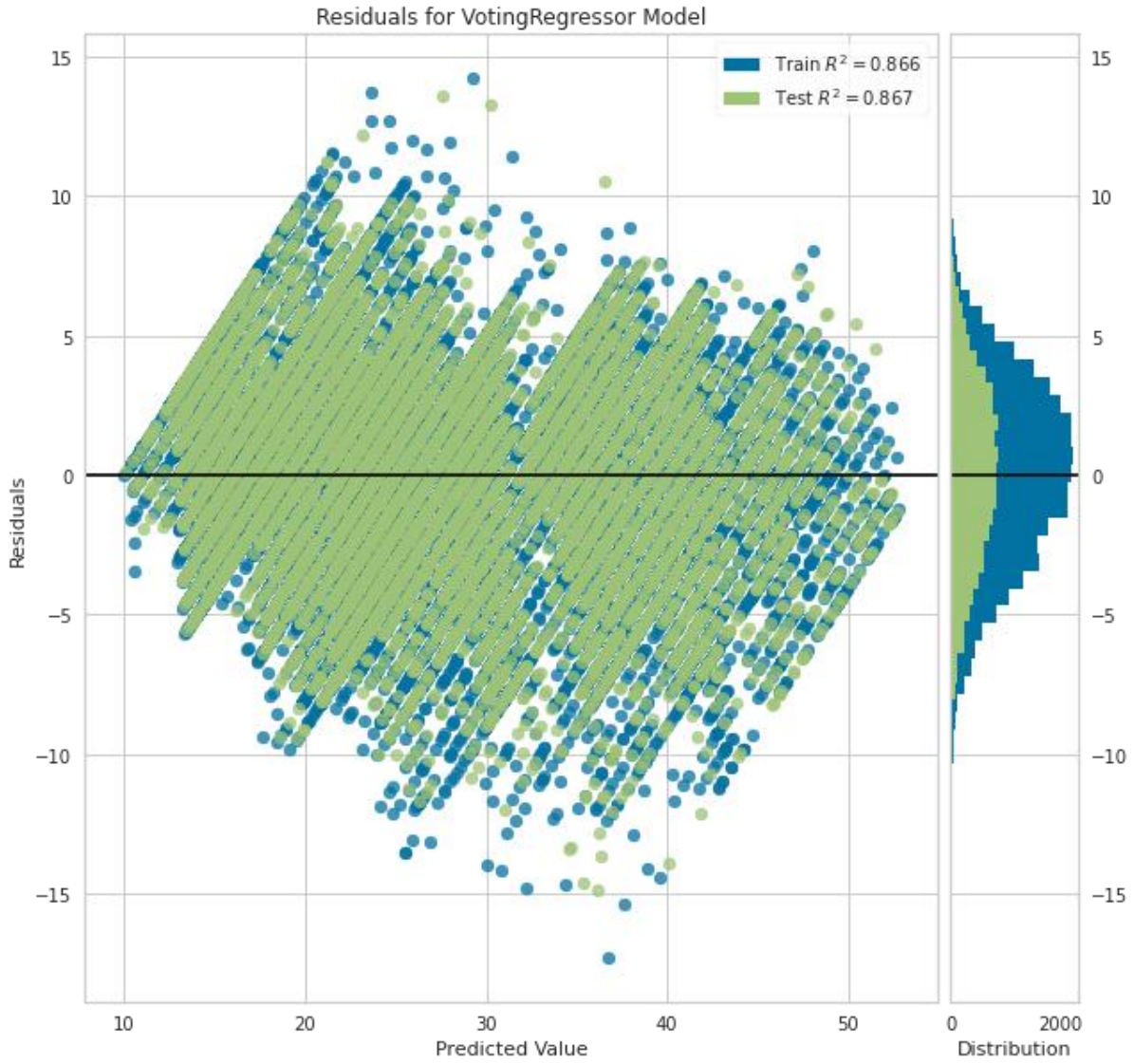
	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
ada	AdaBoost Regressor	4.9763	36.7258	6.0596	0.5828	0.2572	0.2392	62.6840
huber	Huber Regressor	4.9429	38.8241	6.2306	0.5589	0.2518	0.2153	58.8220
en	Elastic Net	6.2469	60.1472	7.7550	0.3170	0.2996	0.2784	0.5620
lasso	Lasso Regression	6.2698	60.6911	7.7899	0.3108	0.3007	0.2793	0.5060
knn	K Neighbors Regressor	6.6192	69.3852	8.3295	0.2120	0.3200	0.2934	0.6600
llar	Lasso Least Angle Regression	7.5890	88.0981	9.3857	-0.0005	0.3611	0.3447	1.0920
dummy	Dummy Regressor	7.5890	88.0981	9.3857	-0.0005	0.3611	0.3447	0.1620
par	Passive Aggressive Regressor	7.5579	92.0244	8.9595	-0.0532	0.4466	0.3117	6.5140
lr	Linear Regression	7.6146	258.9188	13.5609	-1.8902	0.4209	0.3328	0.9260

- Blended the top 3 model

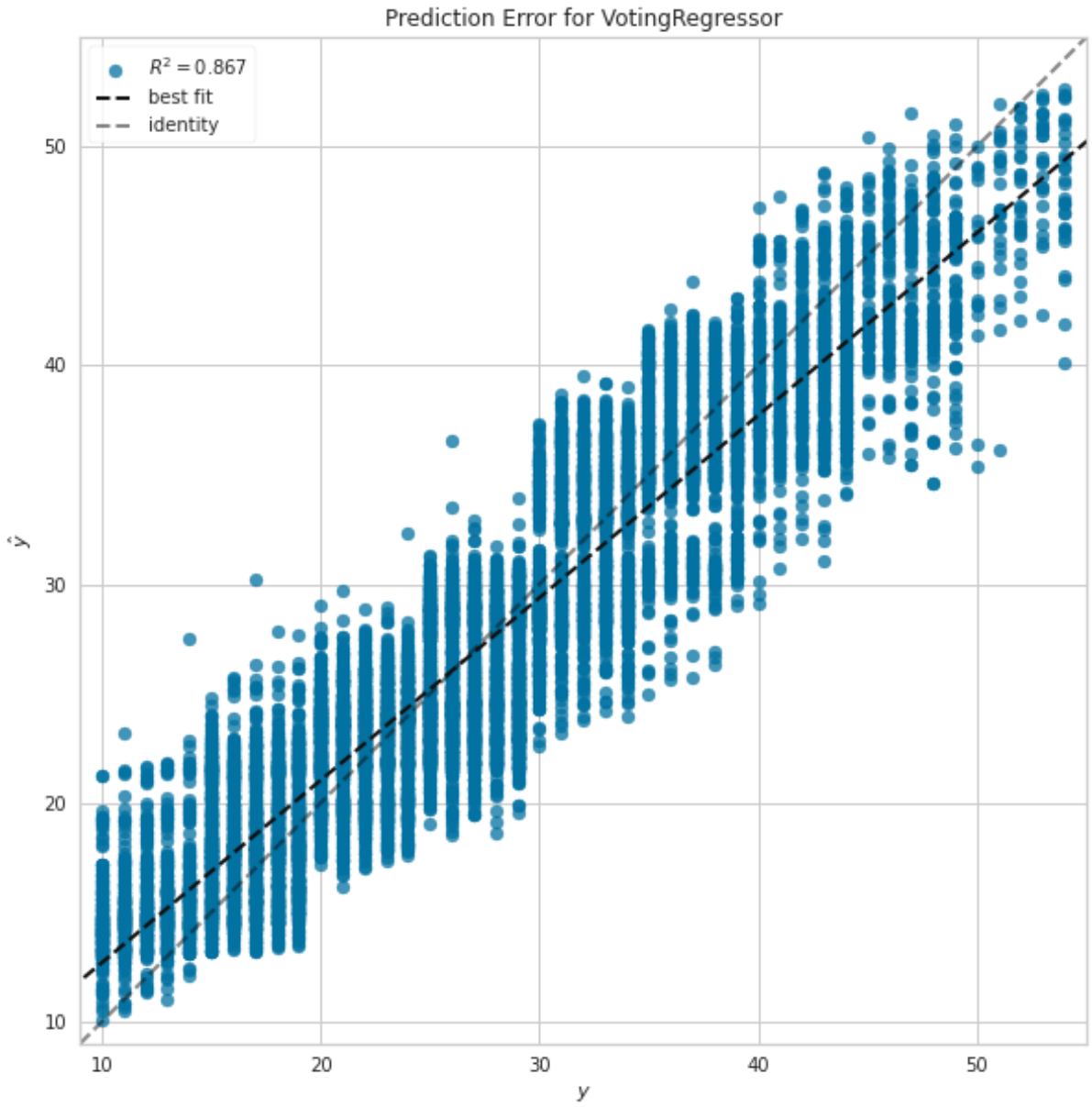


Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	3.1213	15.1358	3.8905	0.8247	0.1615	0.1369
1	3.1275	15.3078	3.9125	0.8260	0.1596	0.1347
2	3.1043	15.4725	3.9335	0.8235	0.1622	0.1362
3	3.1354	15.5502	3.9434	0.8286	0.1606	0.1355
4	3.1051	15.3104	3.9129	0.8252	0.1620	0.1364
Mean	3.1187	15.3554	3.9185	0.8256	0.1612	0.1359
Std	0.0123	0.1443	0.0184	0.0017	0.0009	0.0008

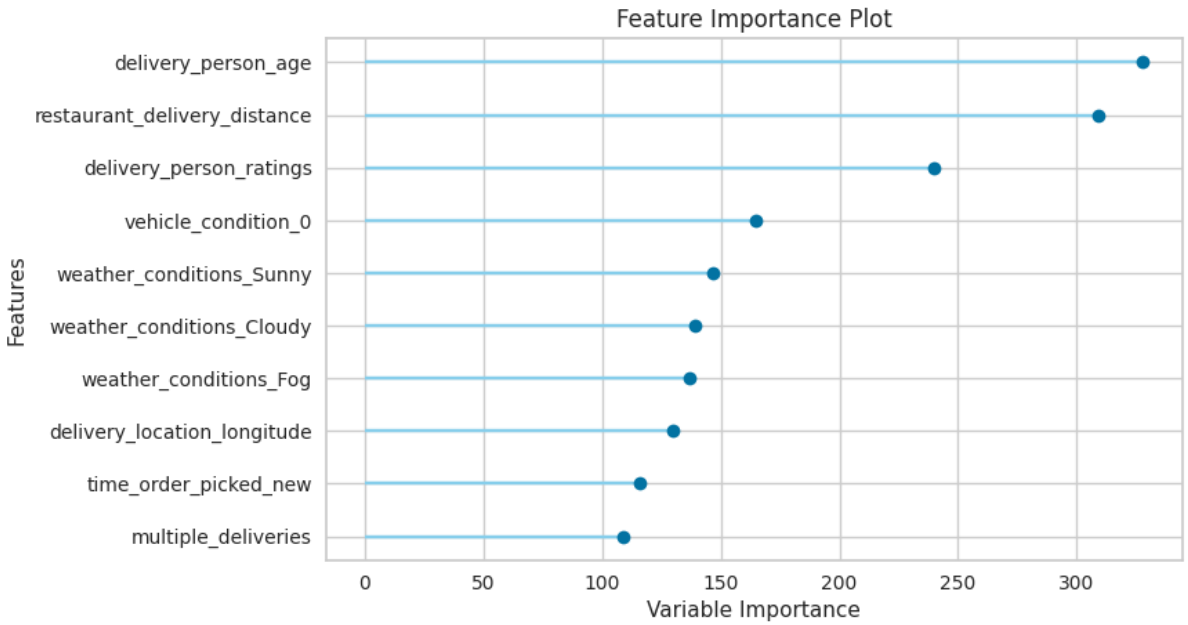
- Lightgbm Regressor Residual Plot



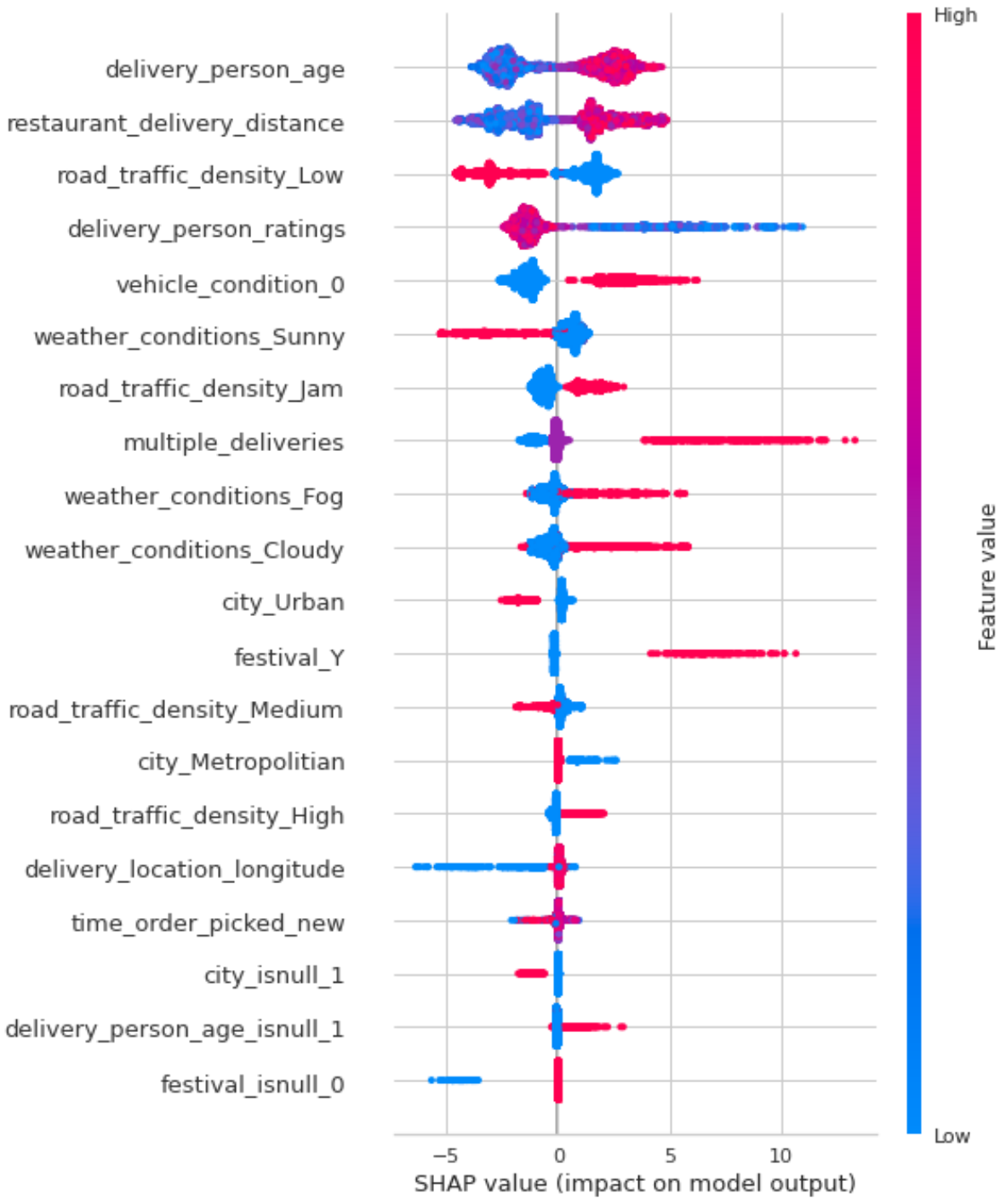
- Lightgbm Regressor Prediction Error Plot



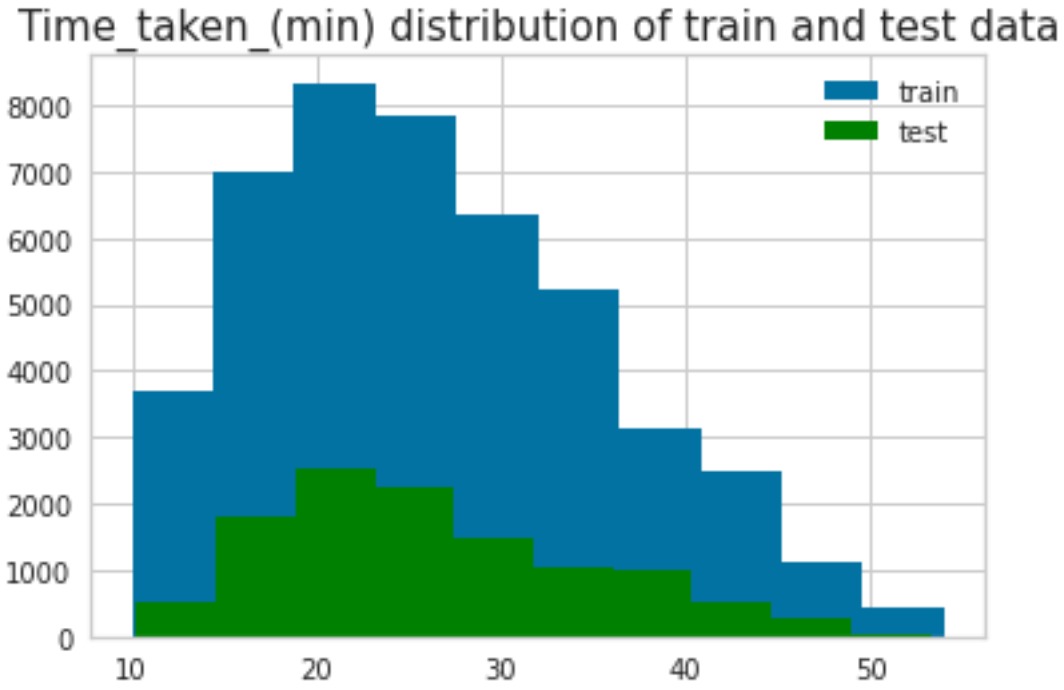
- Lightgbm Model Feature Importance Plot



- SHAP - Lightgbm Model Feature Importance Plot



- Time_taken_(min) distribution of train and test data



- Final score is 82.15