Hackerearth-Get-A-Room-ML-Hackathon-2022

Build a Machine Learning model to identify the habitability score of the property based on the property's basic information and location-based information.

- Basic exploratory data analysis using pandas, matplotlib, seaborn packages.
- Data pre-processing
    - Missing value indicator
    - Missing value imputation for the columns,
        - property_type
        - number_of_windows
        - furnishing
        - frequency_of_powercuts
        - crime_rate
        - dust_and_noise
    - Feature Engineering
        - doors_windows_ratio
        - air_quality_index_category
        - Numerical feature engineering
            - Groupby numerical summary(min,mean, median, max) of numerical columns.

        - z-score outlier indicator for numerical columns.

- The final features for the model
    - 0_property_type
    - 1_property_area
    - 2_number_of_windows
    - 3_number_of_doors
    - 4_furnishing
    - 5_frequency_of_powercuts
    - 6_power_backup
    - 7_water_supply
    - 8_traffic_density_score
    - 9_crime_rate
    - 10_dust_and_noise
    - 11_air_quality_index
    - 12_neighborhood_review
    - 13_property_type_is_null
    - 14_number_of_windows_is_null
    - 15_furnishing_is_null
    - 16_frequency_of_powercuts_is_null
    - 17_crime_rate_is_null
    - 18_dust_and_noise_is_null
    - 19_doors_windows_ratio
    - 20_air_quality_index_category
    - 21_property_area_mean

- 22_property_area_median
- 23_property_area_min
- 24_property_area_max
- 25_number_of_windows_mean
- 26_number_of_windows_median
- 27_number_of_windows_min
- 28_number_of_windows_max
- 29_number_of_doors_mean
- 30_number_of_doors_median
- 31_number_of_doors_min
- 32_number_of_doors_max
- 33_traffic_density_score_mean
- 34_traffic_density_score_median
- 35_traffic_density_score_min
- 36_traffic_density_score_max
- 37_air_quality_index_mean
- 38_air_quality_index_median
- 39_air_quality_index_min
- 40_air_quality_index_max
- 41_neighborhood_review_mean
- 42_neighborhood_review_median
- 43_neighborhood_review_min
- 44_neighborhood_review_max
- 45_property_area_outlier
- 46_traffic_density_score_outlier

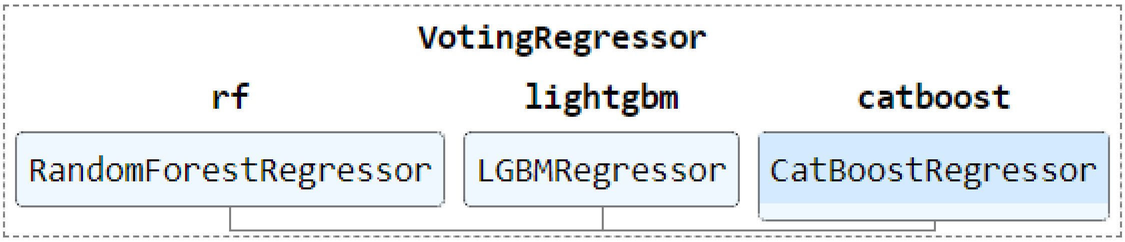
- 47_air_quality_index_outlier
- 48_neighborhood_review_outlier
- 49_habitability_score

- 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)
rf	Random Forest Regressor	4.6017	35.8650	5.9880	0.8210	0.0955	0.0698	5.6140
lightgbm	Light Gradient Boosting Machine	4.8186	37.5961	6.1305	0.8124	0.0962	0.0725	0.7300
catboost	CatBoost Regressor	4.7916	37.5905	6.1305	0.8124	0.0975	0.0725	7.3900
xgboost	Extreme Gradient Boosting	4.8870	39.3572	6.2729	0.8036	0.0994	0.0737	0.8160
et	Extra Trees Regressor	4.8310	40.3693	6.3532	0.7985	0.1004	0.0731	29.5400
gbr	Gradient Boosting Regressor	5.6450	49.7036	7.0492	0.7520	0.1121	0.0858	6.7500
dt	Decision Tree Regressor	6.0773	67.5450	8.2176	0.6629	0.1302	0.0917	0.5040
ada	AdaBoost Regressor	6.7574	71.3201	8.4434	0.6442	0.1425	0.1070	4.7220
br	Bayesian Ridge	7.2505	81.5571	9.0303	0.5931	0.1560	0.1182	0.4820
omp	Orthogonal Matching Pursuit	7.2705	82.0004	9.0547	0.5909	0.1564	0.1186	0.0780
lasso	Lasso Regression	7.6056	103.9907	10.1965	0.4813	0.1855	0.1338	0.2040
en	Elastic Net	8.1453	129.1196	11.3619	0.3560	0.2082	0.1490	0.0540
huber	Huber Regressor	8.6962	147.6795	12.1500	0.2626	0.2186	0.1601	6.5320
llar	Lasso Least Angle Regression	9.8303	200.4951	14.1587	-0.0001	0.2486	0.1840	0.0780

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
dummy	Dummy Regressor	9.8303	200.4951	14.1587	- 0.0001	0.2486	0.1840	0.0160
ridge	Ridge Regression	10.9717	218.4181	13.7924	- 0.0877	0.2326	0.1674	0.0340
knn	K Neighbors Regressor	10.6929	220.3862	14.8446	- 0.0994	0.2546	0.1942	0.2220
par	Passive Aggressive Regressor	11.1260	259.7278	16.0460	- 0.2986	0.2715	0.2082	0.6500
lr	Linear Regression	18.2003	727.8763	24.1615	- 2.6292	0.4049	0.2753	0.0340
lar	Least Angle Regression	833.0000	8388.0	4066.00	404.00	26.9127	11.00	0.1460

- Blended the top 3 model



VotingRegressor

rf

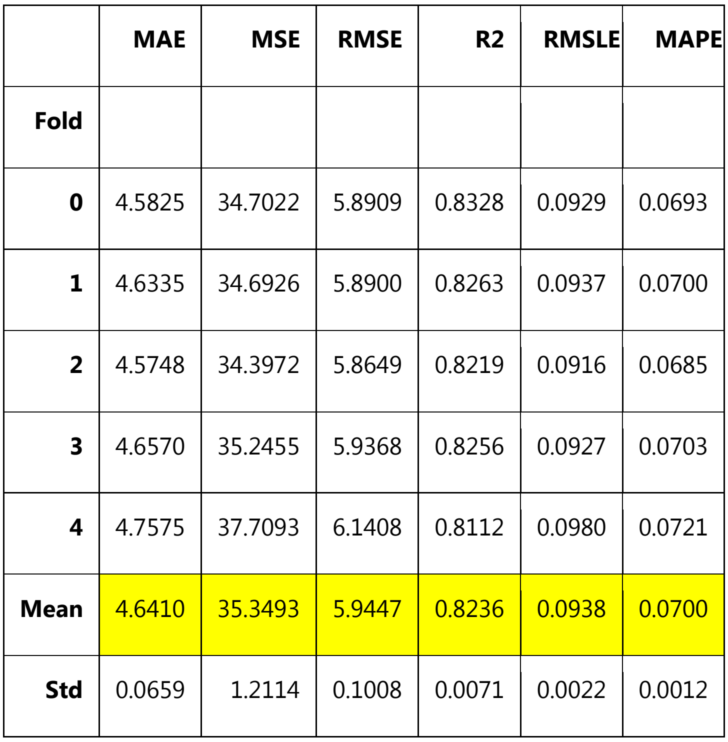
RandomForestRegressor

lightgbm

LGBMRegressor

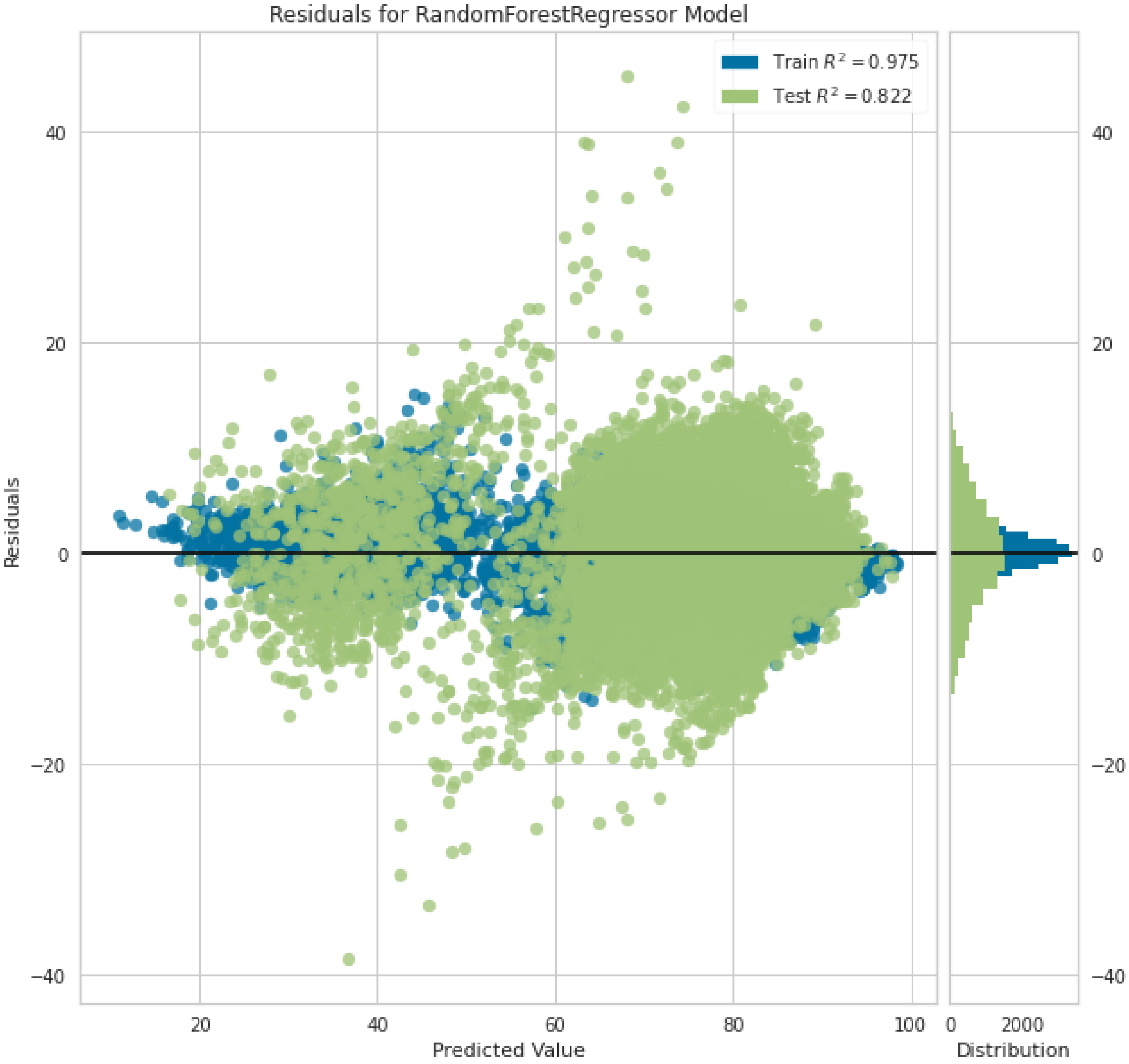
catboost

CatBoostRegressor

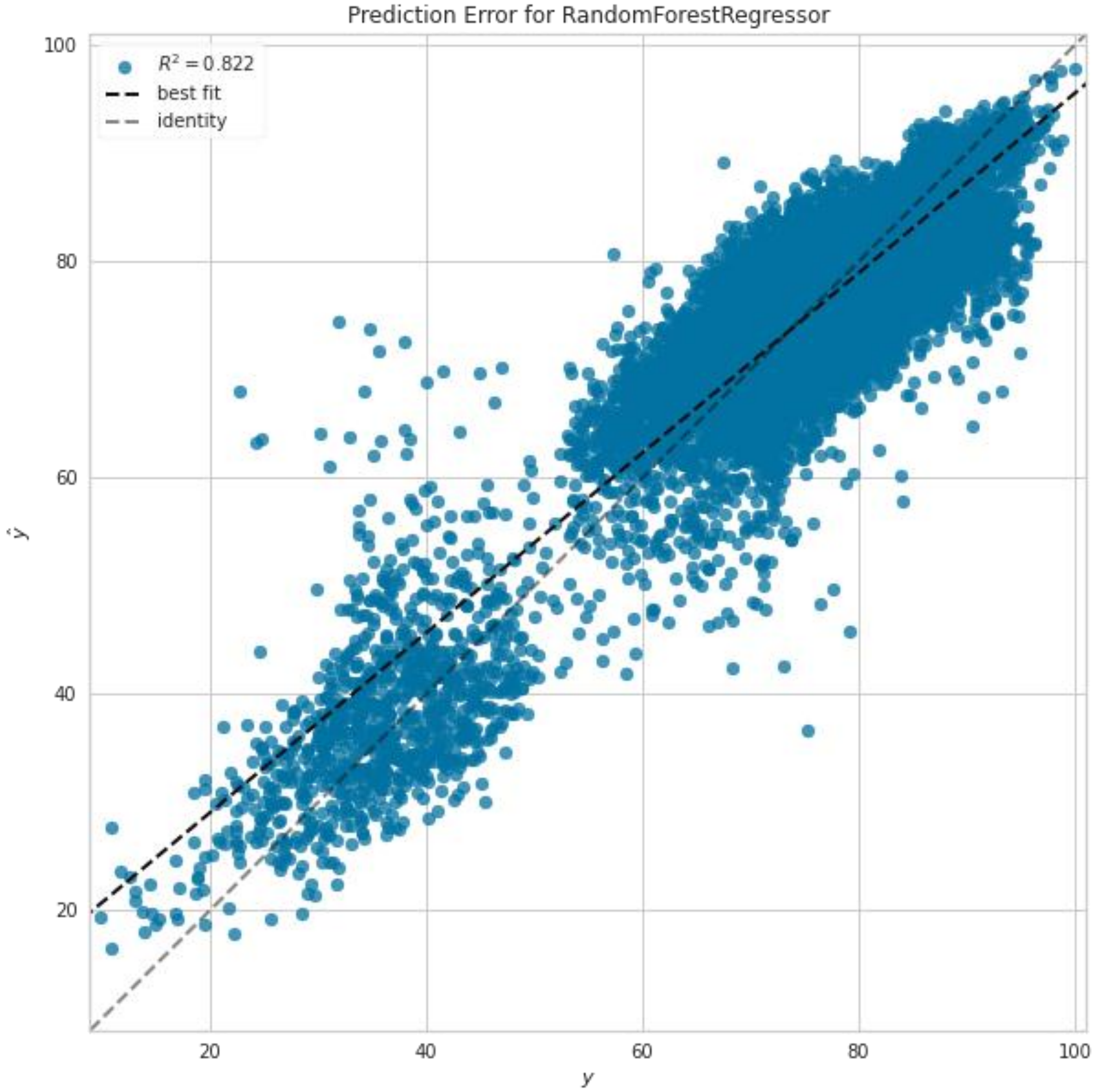


	MAE	MSE	RMSE	R2	RMSLE	MAPE
Fold						
0	4.5825	34.7022	5.8909	0.8328	0.0929	0.0693
1	4.6335	34.6926	5.8900	0.8263	0.0937	0.0700
2	4.5748	34.3972	5.8649	0.8219	0.0916	0.0685
3	4.6570	35.2455	5.9368	0.8256	0.0927	0.0703
4	4.7575	37.7093	6.1408	0.8112	0.0980	0.0721
Mean	4.6410	35.3493	5.9447	0.8236	0.0938	0.0700
Std	0.0659	1.2114	0.1008	0.0071	0.0022	0.0012

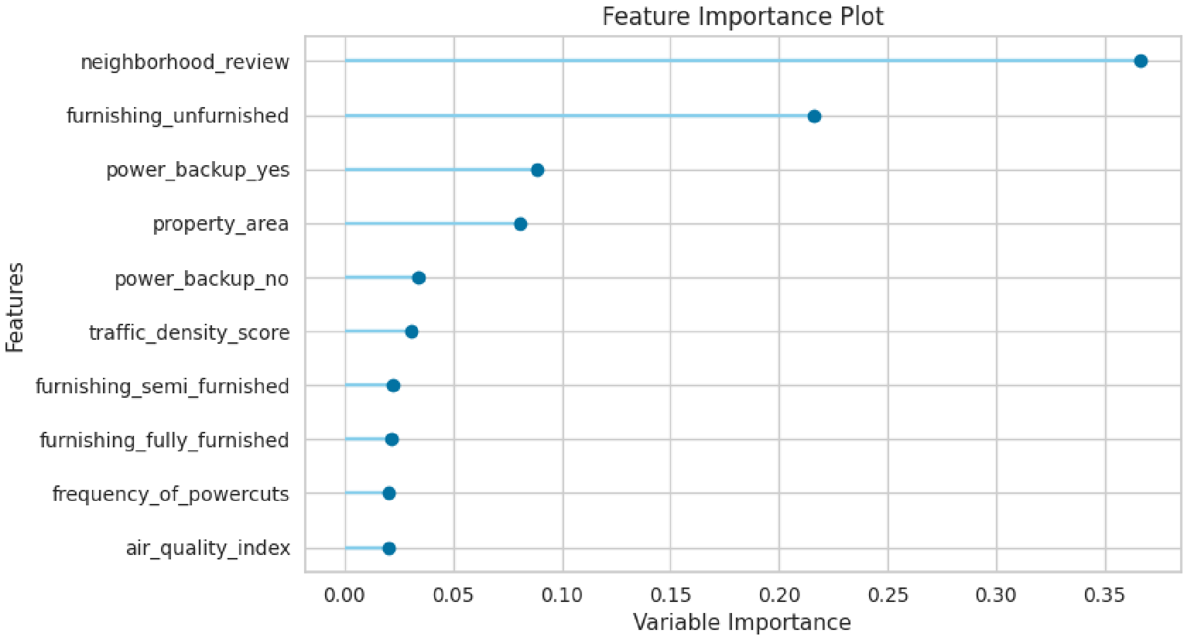
- Random Forest Regressor Residual Plot



- Random Forest Regressor Prediction Error Plot



- Random Forest Model Feature Importance Plot


Feature Importance Plot
Features
neighborhood_review
furnishing_unfurnished
power_backup_yes
property_area
power_backup_no
traffic_density_score
furnishing_semi_furnished
furnishing_fully_furnished
frequency_of_powercuts
air_quality_index
Variable Importance