Prediction Results

1 Regression for each Area with area related features

The model was trained on areas with more than 40000 records, resulting in 5 different areas. To get a better understanding of how good or bad the error metrics are, the mean value and variation for each area is also presented. Note that anomaly samples are removed before training.

Area: Marsa Dubai

mean	16424.17
variation	29901.51

Regressor	MAE	RMSE
Linear Regression	2944.49	2944.49
XGBoost	3007.04	3775.76
Stochastic Gradient Descent	2982.43	3759.08
Random Forest	2936.1	3731.74
Neural Network (MLP)		

Area: Al Thanyah Fifth

mean	9367.46
variation	17808.6

Regressor	MAE	RMSE
Linear Regression	11 digits!	11 digits!
XGBoost	2161.96	2916.93
Stochastic Gradient Descent	2101.81	2864.84
Random Forest	2125.56	2873.81
Neural Network (MLP)		

Area: Al Warsan Firsth

mean	7155.52
variation	53744.9

Regressor	MAE	RMSE

Linear Regression	13 digits!	13 digits!
XGBoost	1486.49	1952.4
Stochastic Gradient Descent	1415.55	1878.32
Random Forest	1333.12	1804.41
Neural Network (MLP)		

Area: Business Bay

mean	16941.93
variation	41930.07

Regressor	MAE	RMSE
Linear Regression	8 digits!	8 digits!
XGBoost	2537.67	3302.45
Stochastic Gradient Descent	2522.94	3305.88
Random Forest	2502.86	3303.7
Neural Network (MLP)		

Area: Al Barsha South Fourth

mean	10022.98
variation	38542.77

Regressor	MAE	RMSE
Linear Regression	1813.41	2599.18
XGBoost	1895.63	1895.63
Stochastic Gradient Descent	1854.5	2596.6
Random Forest	1803.27	2579.7
Neural Network (MLP)		

I repeated the training phase with two other scenarios. One with the unit related features and the other with the complete feature set. In none of the scenarios the errors had less than 4 digits.

By far, among the first four regression models, random forest is doing better. But an optimal subset of features must be selected. In the following days, I try to work around features to see how they impact the performance of the models.