Springboard Capstone Project 1
Nancy Mao | dongxiunan.mao@gmail.com

Beijing House Pricing Prediction

In Depth Analysis

Goal

In this section, we will build a regression model to predict the housing price. We will try out several regression models and then select the model with lowest RMSE as our model and do hyperparameter tuning.

Feature Engineering

- New features created:
    - 'FloorNumber' - Indicates which floor the house is at. Estimated based on floorPosition and buildingFloors.
- Feature selection
    - 19 features are included to build the model
        - 'Lng', 'Lat', 'tradeTime', 'DOM', 'followers','bedRoom', 'livingRoom', 'kitchen', 'bathRoom', 'buildingType', 'constructionTime', 'renovationCondition', 'buildingStructure', 'elevator', 'fiveYearsProperty', 'subway', 'district', 'communityAverage', 'floorNumber'

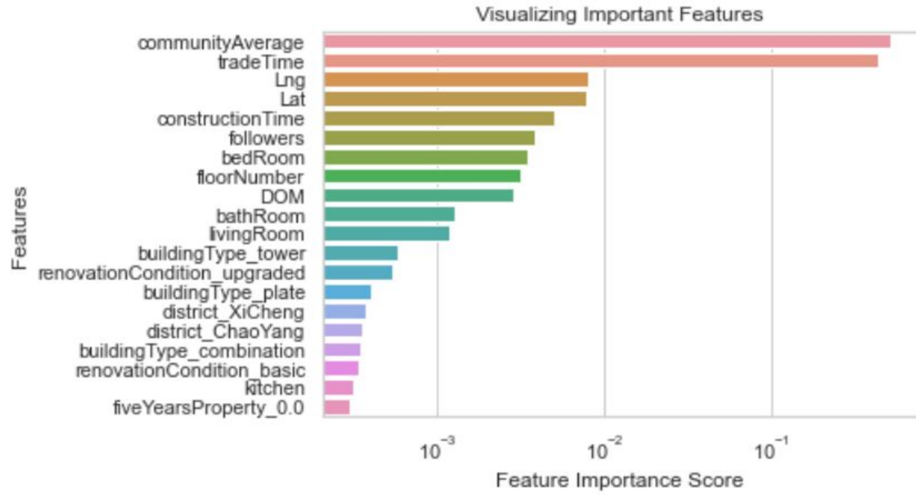
Model Comparison and Performance

Given this is my first project I decided to start with six classic machine learning regressors (Linear Regression, Ridge Regression, Lasso Regression, Linear SVR, Random Forest and Gradient Boosting).

	model	train_RMSE	test_RMSE
0	LR	8944.587	8962.582
1	RIDGE	8941.056	8958.310
2	LASSO	22216.348	22037.518
3	LSVR	10765.068	10709.196
4	RF	3710.491	5204.277
5	GB	1338.953	5150.458

Random Forest and Gradient Boosting have better performance in terms of low RMSE but Gradient Boosting tends to overfit more than Random Forest does. I decided to choose Random

Forest as my regression model and move forward with hyperparameter tuning to see if the performance can be improved.



Hyperparameter Tuning - Random Forest

Tuning Results

Parameter	Best Params
'max_depth'	40
'min_samples_leaf'	2
'n_estimators'	100

Model Performance after Tuning:

- Test_Rmse: 5103.172996944146