Analzye and predict house price in King County, Seattle

Chenghua Wu

Abstract

21613 records of house Sales in King County, USA which includes Seattle is used to determine key features for house price by plotting and predict house price by machine learning models. Based on plotting, Bathroom, bedroom , sqft_above, sqft_basement, sqft_living15, sqft_lot15 are the key features that determine house price. When using LinearRegression, the whole dataset is more accurate than hand-picked key features (r2 score = 0.698 vs 0.519). When using KNeighborsRegressor, key features dataset is more accurate than the whole dataset (r2 score = 0.536 vs 0.483)

Motivation

What are the key features that determines house price?

How to predict hourse price?

Dataset(s)

- House Sales in King County, USA which includes Seattle. It includes homes sold between May 2014 and May 2015.
- The dataset has 21613 records
- Download from https://www.kaggle.com/harlfoxem/housesalesprediction

Data Preparation and Cleaning

As I checked, there is no missing data. Dataset is very clean and ready to be analyzed.

Research Question(s)

- 1.What are the key features to determine or predict house prices?

- 2.Given the house prices varied widely, Divide the house price into three categories: top 10% (25%), bottom 10%(25%), the rest in the middle, to see if key features were changed for house prices?

- 3.Try different Machine Learning model to predict house prices. Linear Regression or KNN regression, which one is more accurate?

Methods

1. Use seaborn to visualize data

2. Use Linear Regression and KNeighborsRegressor from sklearn to predict house prices.

House price summary

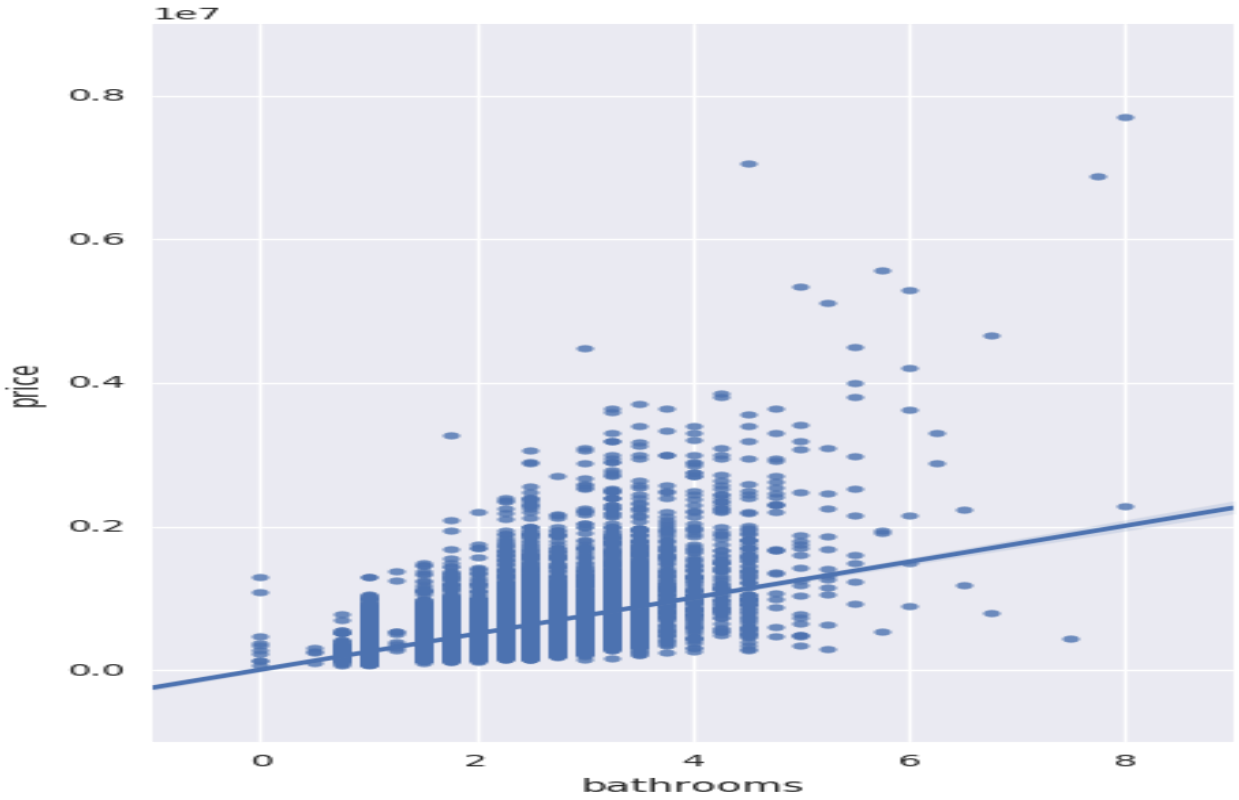
mean	std	min	25%	50%	75%	max
540088.14	367127.20	75000.00	321950.00	450000.00	645000.00	7700000.00

Mean of house price is 540k; Price varied greatly, the highest price is 100 times of the lowest price

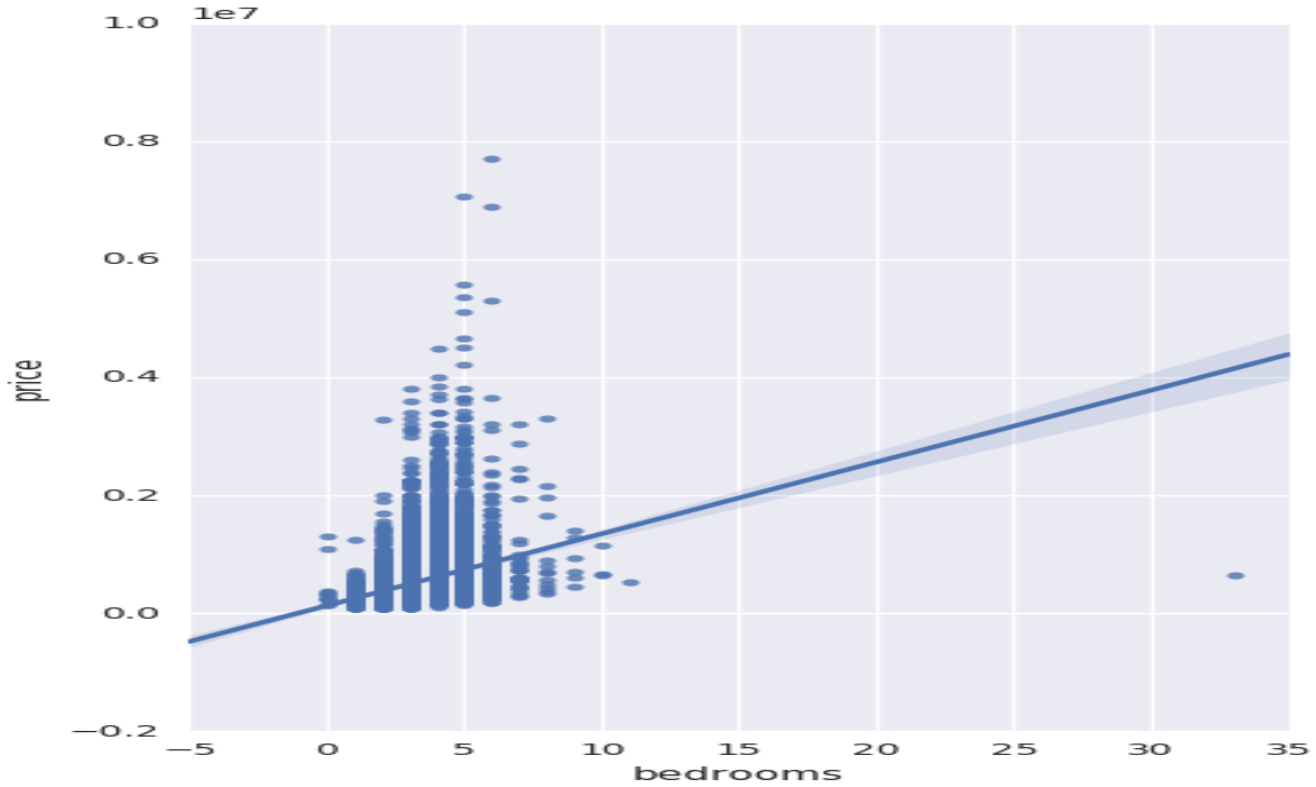
Plot features vs price

Bathroom, bedroom , sqft_above, sqft_basement, sqft_living15, sqft_lot15 are relatively correlated with price by judging the slope of regression line and the scatter dot.

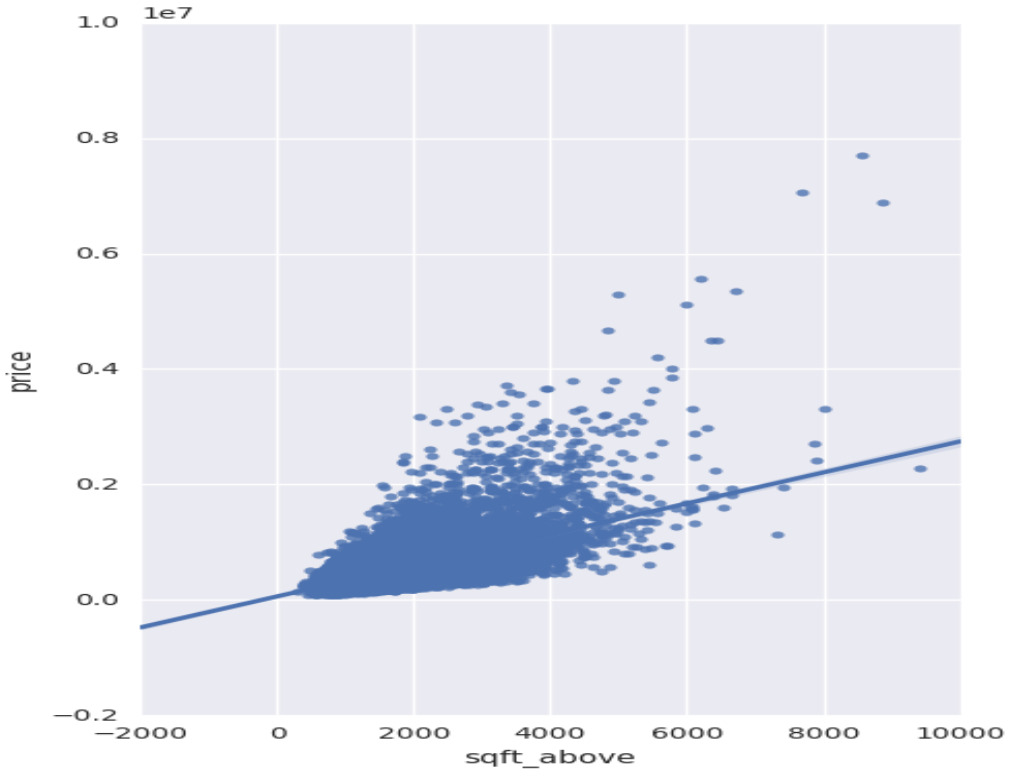
Bathrooms vs price



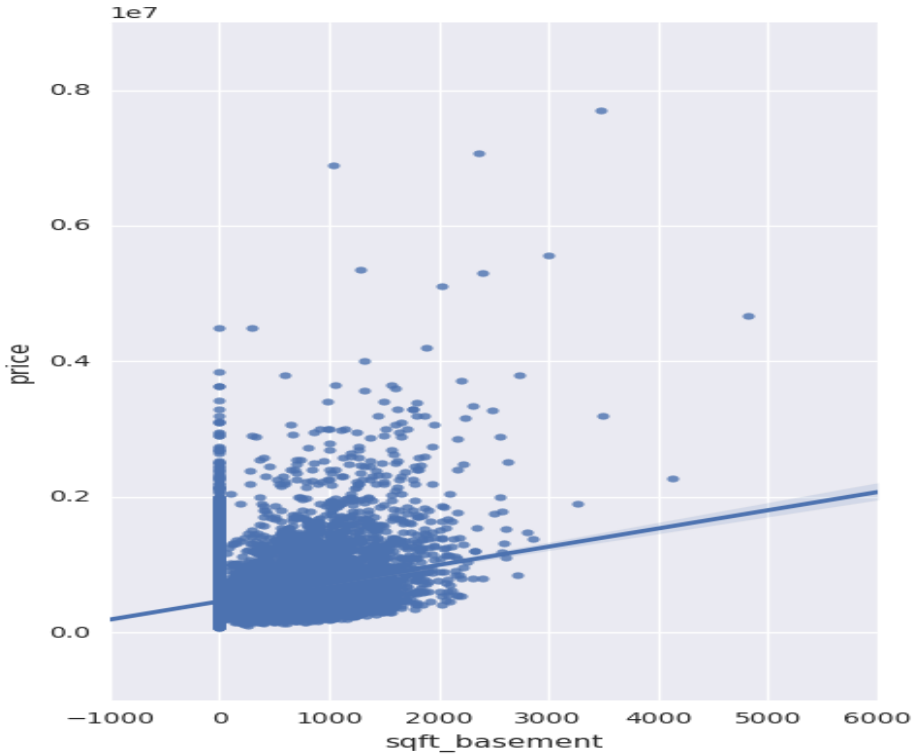
Bedroom vs price



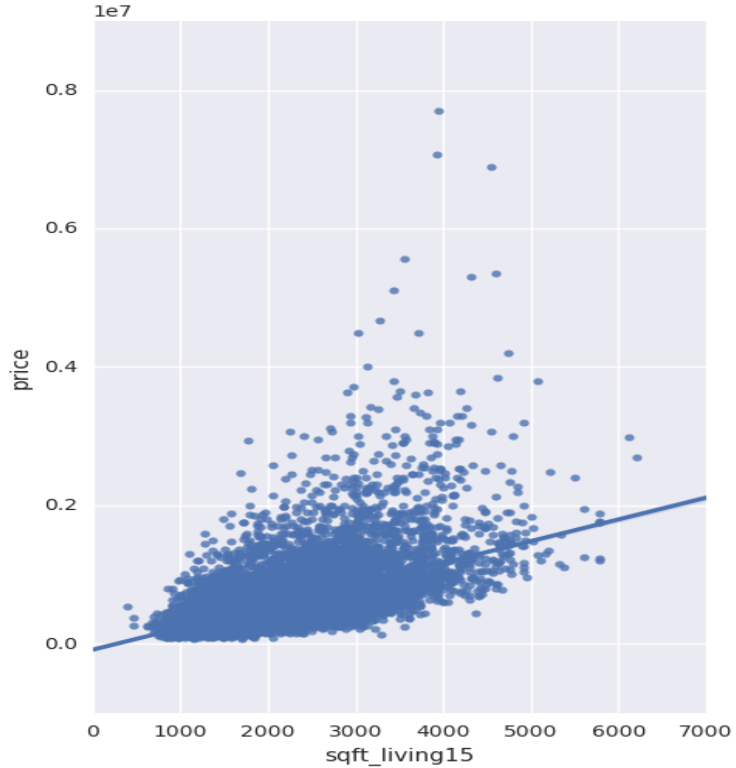
Sqft_above vs price



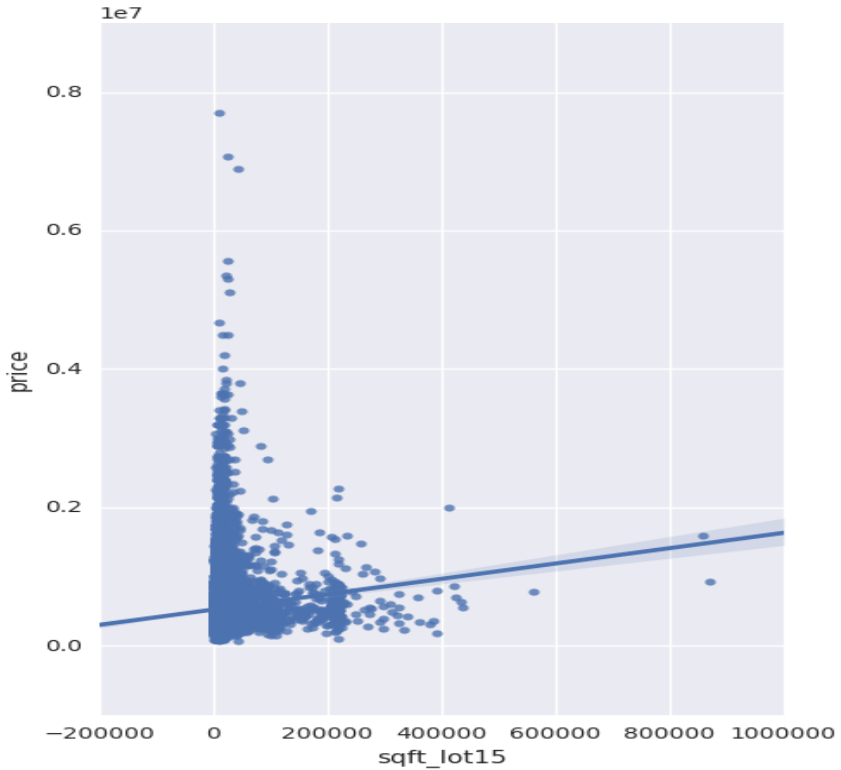
sqft_basement vs price



Sqft_living15 vs price



sqft_lot15 vs price



Divide dataset into bottom 25%, middle, top 25% based on price

Plot 3 datasets, they are showing similar regression lines to the whole dataset.

For each dataset, regression line and data scattering points are varied from each dataset. This means that 3 datasets share key features in term of determing house price, but the key features are varied from dataset to dataset to determine house price.

Charts are not shown here, but it is in the notebook

Median Price for the bottom and top 10 zip code

Median Price for the bottom 10 zip code	Median Price for the top10 zip code
zipcode price 98002 235000.00 98168 235000.00 98032 249000.00 98001 260000.00 98188 264000.00 98198 265000.00 98003 267475.00 98023 268450.00 98148 278000.00 98178 278277.00	zipcode price 98102 720000.00 98109 736000.00 98075 739999.00 98119 744975.00 98006 760184.50 98005 765475.00 98112 915000.00 98040 993750.00 98004 1150000.00 98039 1892500.00

Predict house price based on key features

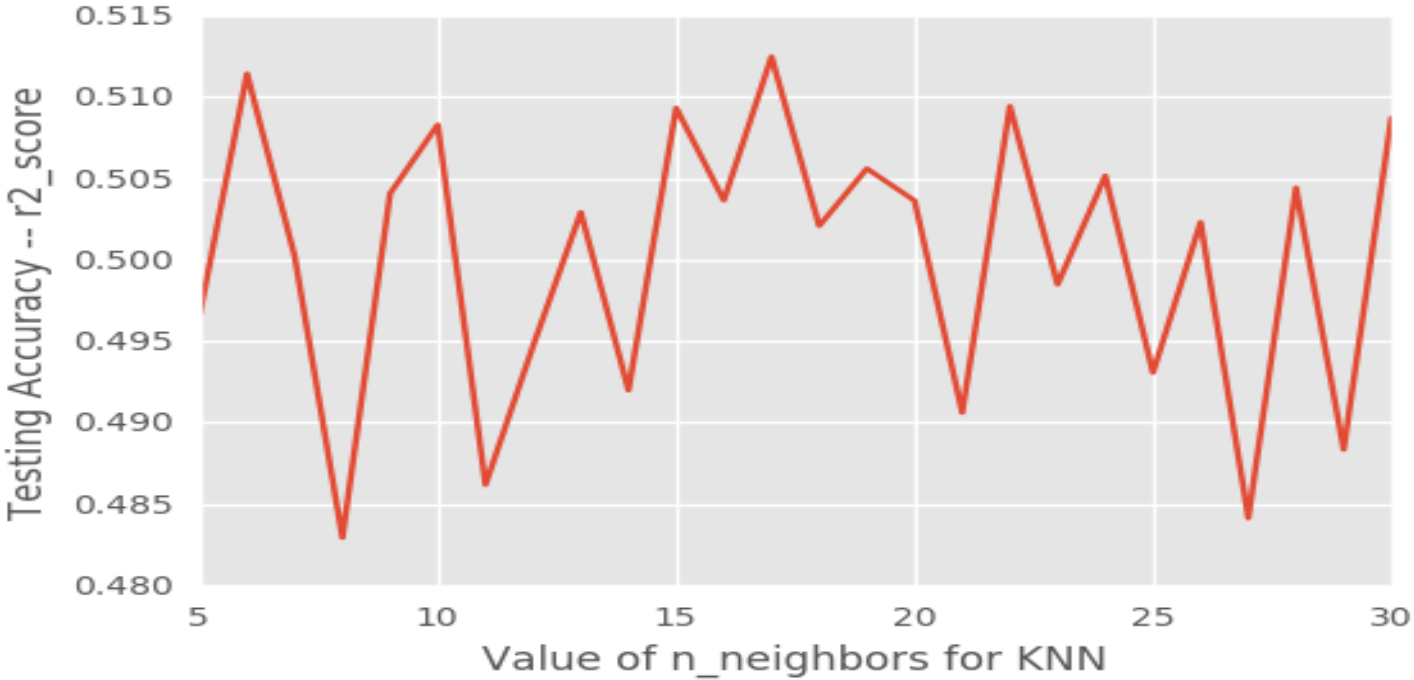
For key features: 'bedrooms', 'bathrooms', 'sqft_above', 'sqft_basement', 'sqft_living15', 'sqft_lot15‘, use LinearRegression and KNeighborsRegressor to predict house price

LinearRegression, r2 score = 0.519

KNeighborsRegressor, r2 score = 0.536 (*n_neighbors=5*)

Evaluate n_neighbors for KNN in key features dataset

.



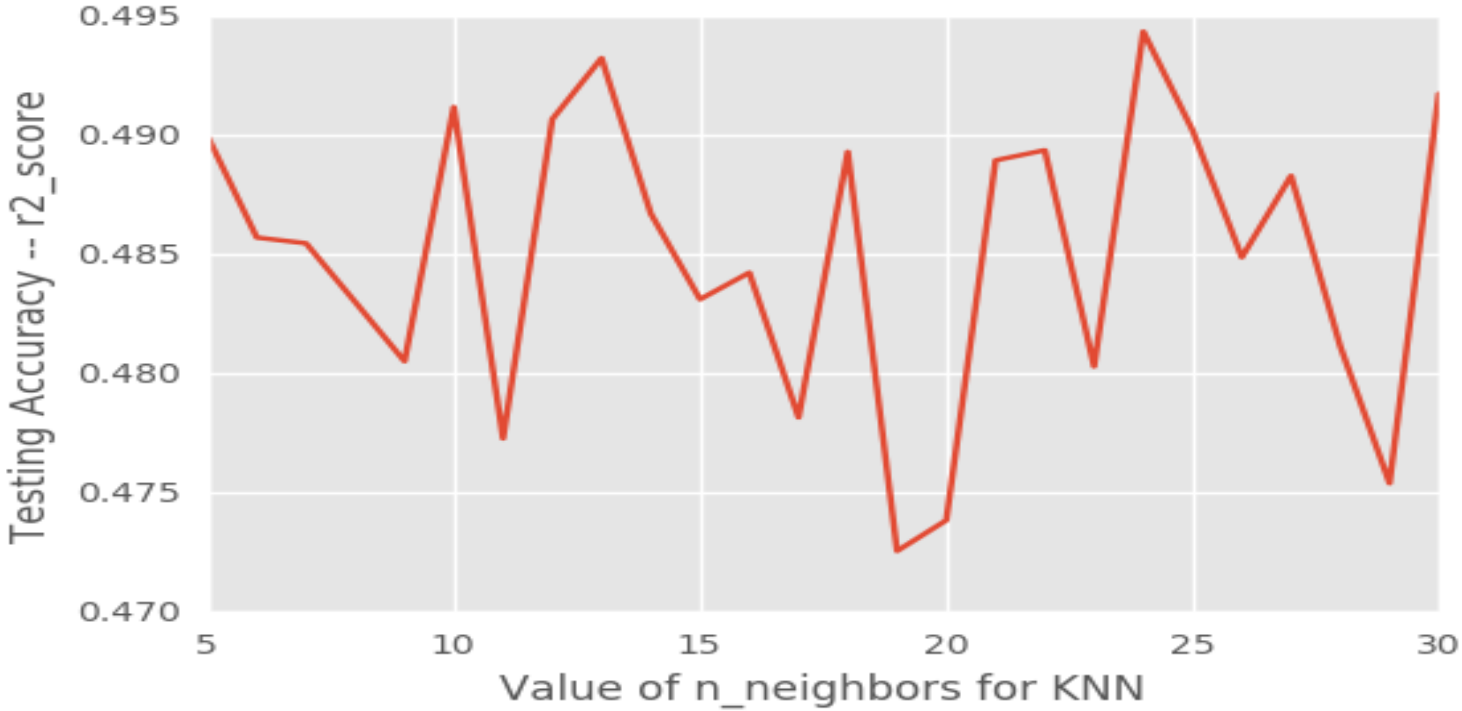
Predict house price based on all features

Use all features to predict house price

LinearRegression, r2 score = 0.698

KNeighborsRegressor, r2 score = 0.483 (*n_neighbors=5*)

Evaluate n_neighbors for KNN in the whole dataset



Conclusions

Based on plot, Bathroom, bedroom , sqft_above, sqft_basement, sqft_living15, sqft_lot15 are the key features that determine house price.

When using LinearRegression, the whole dataset is more accurate than hand-picked key features (r2 score = 0.698 vs 0.519)

When using KNeighborsRegressor, key features dataset is more accurate than the whole dataset (r2 score = 0.536 vs 0.483)

Therefore, features should not be discarded easily; Dependent on different models, it may have different predicting result.

Limitations and future plan

Limitation:

School district is highly affected house price. This information is not in the dataset

Future plan:

1. Use machine learning models to select features
2. Try more machine models to predict house price

References

Linear Regression:

https://github.com/justmarkham/DAT8/blob/master/notebooks/10_linear_regression.ipynb

Cross-validation for parameter tuning, model selection, and feature selection:

https://github.com/justmarkham/scikit-learn-videos/blob/master/07_cross_validation.ipynb