Machine Learning Engineer Nanodegree Capstone Project - Seattle House Grade Forecast

Lingyun Hu, Nov 2018

Definition

This project will employ several supervised learning algorithms to forecast the house grade in Seattle. The goal is to construct a model that accurately predicts house grade based on the 19 feature columns. The best candidate algorithm will be chosen from preliminary results and will be further optimized based on the historical house transaction data.

The target measure is the overall grade given to the housing unit, based on King County grading system. It is whole number varying from 1 to 10.

Problem Statement

I live in Seattle and am pretty interested in the following questions

- What are the key features determining house grade in Seattle?
- What are the busiest times of the year to buy/sell house in Seattle? By how much more do transactions spike?
- Can I predict house grade based on the given 19 features?

I will use Kc_house_data.csv from Kaggle. It contains following 19 house features along with 21613 observations from 2014 May to 2015 May. I am looking for data after 2015 that can be used to verify my model.

- id: a notation for a house
- date: Date house was sold
- price
- bedrooms
- bathrooms
- sqft_living
- sqft_lot
- floors
- waterfront
- view: Has been viewed
- condition
- sqft_above
- sqft_basement
- yr_built
- yr_renovated

- zipcode
- lat
- long
- sqft_living15: Living room area in 2015(implies-- some renovations) This might or might not have affected the lotsize area
- sqft_lot15: lotSize area in 2015(implies-- some renovations)
- grade

The grade varies from 1 to 13. Based on the definition, score over 10 are errors. I will remove them from the analysis.

Metrics

I will use R2 Score (coefficient of determination) to measure the accuracy of the classifier and evaluate the performance. Best possible R2 score is 1.0 and it can be negative (because the model can be arbitrarily worse). A constant model that always predicts the expected value of y, disregarding the input features, would get a R^2 score of 0.0.

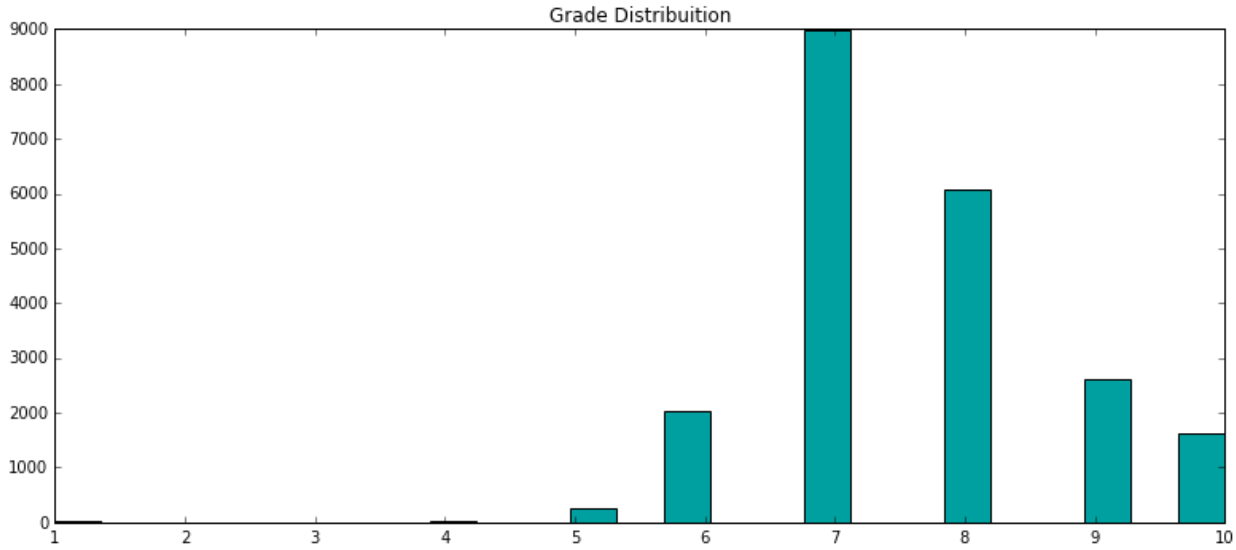
Typical supervised learning models can solve this problem. Some benchmark models are:

- Gaussian Naive Bayes (GaussianNB)
- Decision Trees
- Ensemble Methods (Bagging, AdaBoost, Random Forest, Gradient Boosting)
    - **Extreme Gradient Boosting**: XGBoost is one of the implementations of Gradient Boosting concept, but what makes XGBoost unique is that it uses "a more regularized model formalization to control over-fitting, which gives it better performance," [1]. Therefore, it helps to reduce overfitting.
    - **Random Forest** is a statistical algorithm that is used to cluster points of data in functional groups. When the data set is large and/or there are many variables it becomes difficult to cluster the data because not all variables can be taken into account, therefore the algorithm can also give a certain chance that a data point belongs in a certain group. [2] This is how the clustering takes place.
        - Prepare training set.
        - The algorithm clusters the data into decision tree.
            - At each split or node in this cluster/tree/dendrogram variables are chosen at random by the program to judge whether datapoints have a close relationship or not.
        - The program makes multiple trees a.k.a. a forest. Each tree is different because for each split in a tree, variables are chosen at random.
        - Then the rest of the dataset (not the training set) is used to predict which tree in the forests makes the best classification of the datapoints (in the dataset the right classification is known).
        - The tree with the most predictive power is shown as output by the algorithm.
- K-Nearest Neighbors (KNeighbors)
- Stochastic Gradient Descent Classifier (SGDC)
- Support Vector Machines (SVM)
- Logistic Regression

Analysis

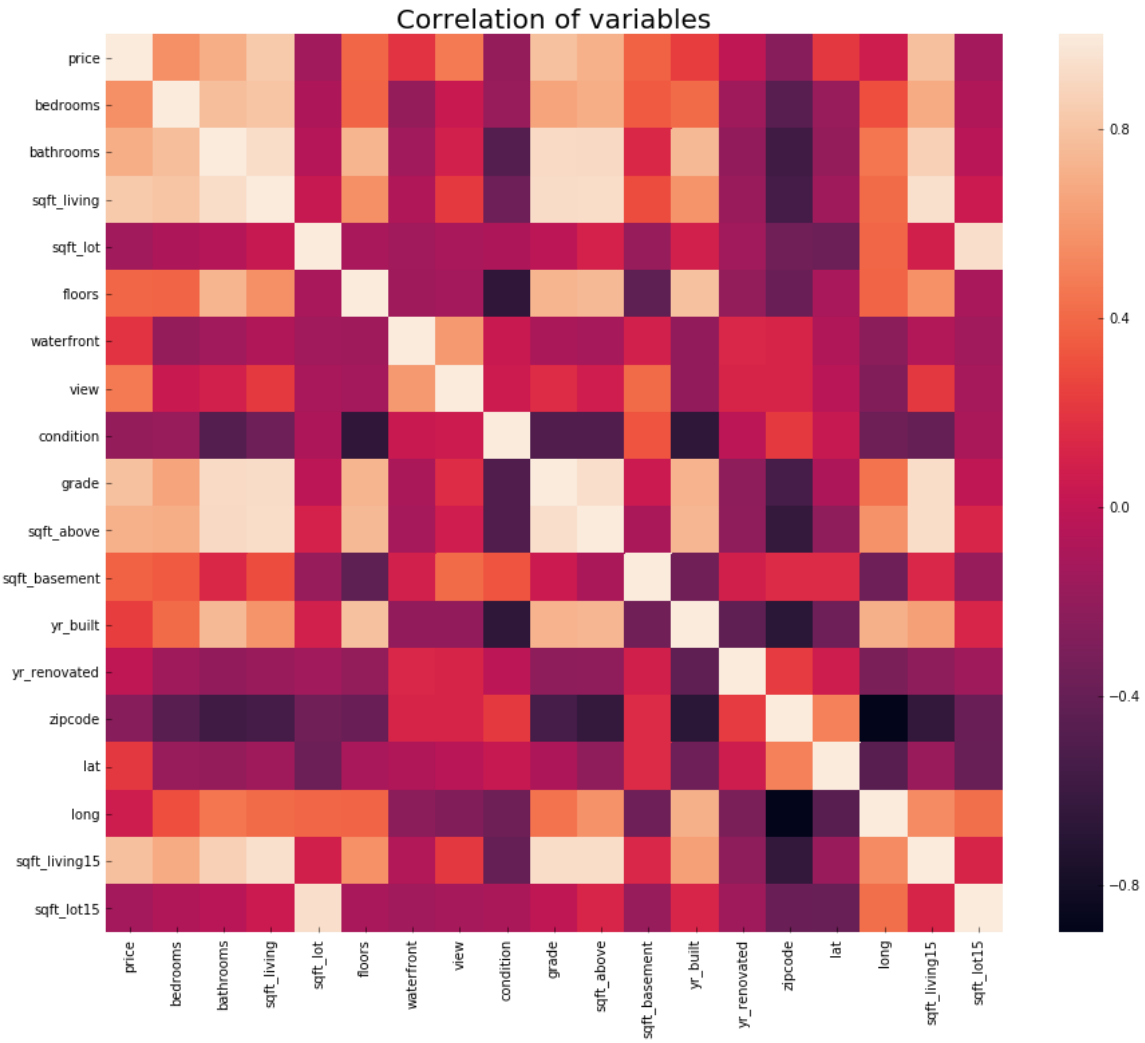
Explore the data

First of all, I will take a look of the grade distribution. There are 21613 records in the file. The mean of grade is 7.6. with std 1.1. It ranges from 1 to 10. 50% of the distribution is 7.



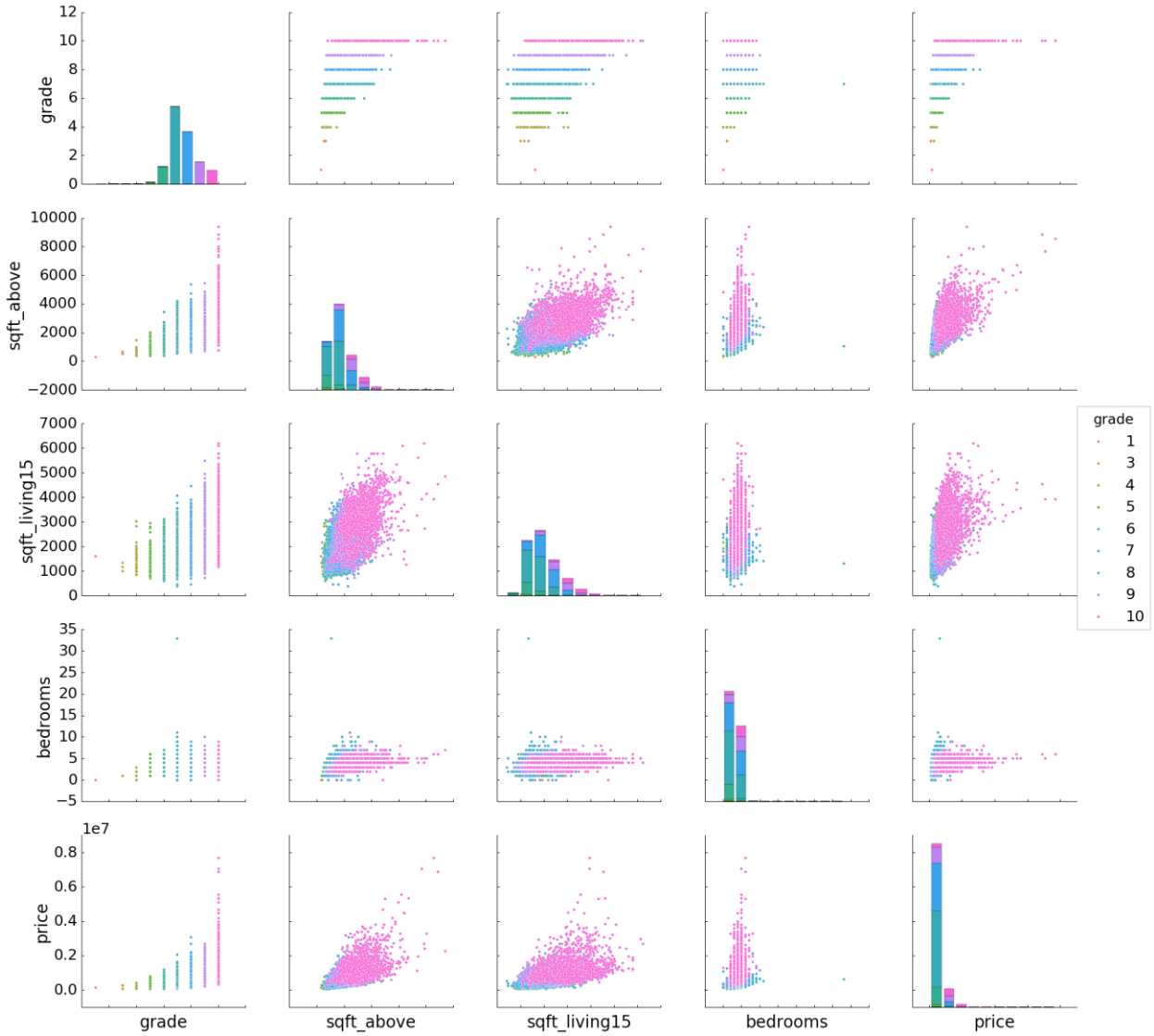
From the Grade Distribution chart, we can see Average grade is 7 and about half of the house got grade of 7.

Secondly, I will check the heat map to get an overall understanding of the relationship between features.



From the heat map result, we can tell the major features those would impact grade include: sqft_above, sqft_living15, sqft_living, bathrooms, price, bedrooms, yr_built,floor. The most unrelated features are condition, zipcode, yr_renovated, lat, waterfall, sqft_lot15, sqft_lot, view, sqft_basement, long.

I try the pair plot to check the relationship among grade and those top 4 impactful features (sqft_above, sqft_living15, bathrooms, price). I am considering sqft_living15 and sqft_living represent similar info here.



What are the key features determining house grade in Seattle?

From the pair plot, we can see low score houses have small lot and living space. and accordingly, less bedrooms and lower prices. Now I will study the rest 3 features (bedrooms, yr_built, floor) in the impactful feature list.

I found from the relationship between grade and bathrooms, floor and building year:

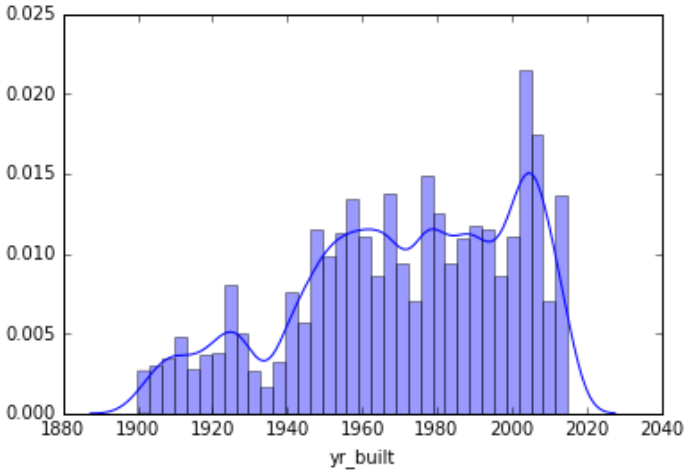
- Grade less than 8 is more likely to have 1 bathroom while grade greater than 7 is more likely to have 2.5 bathrooms. Houses with grade 10 are also very likely to have 3.5 bathrooms. By calculating the bedroom number, we can have a rough idea of the house grade.

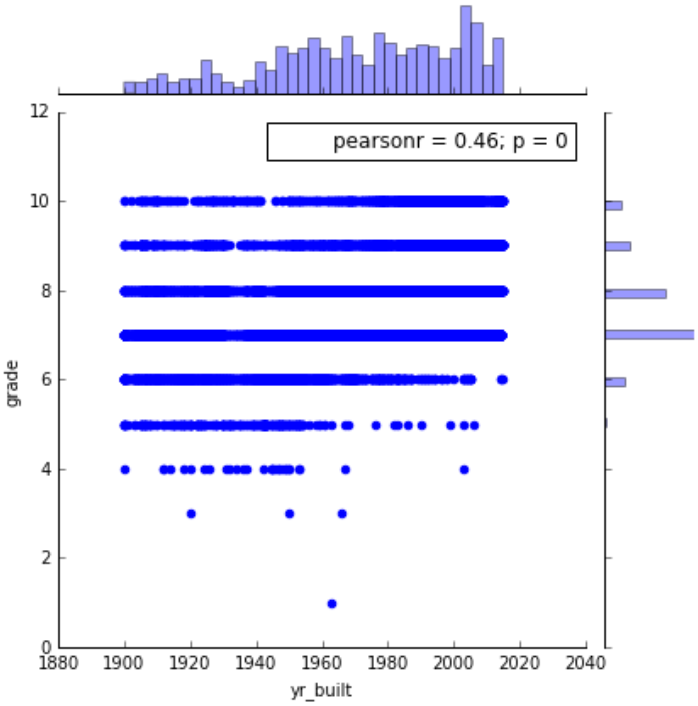
grade	1	3	4	5	6	7	8	9	10
bathrooms									
0.0	1	2	0	0	0	4	2	0	1
0.5	0	0	0	1	2	0	1	0	0
0.75	0	1	14	14	26	17	0	0	0
1.0	0	0	14	190	1414	2084	143	7	0
1.25	0	0	0	0	1	3	2	2	1
1.5	0	0	0	9	137	984	288	23	5
1.75	0	0	1	9	225	1899	808	92	14
2.0	0	0	0	17	184	1165	458	93	13
2.25	0	0	0	0	4	778	956	237	72
2.5	0	0	0	1	24	1390	2278	1226	461
2.75	0	0	0	1	10	322	448	286	118
3.0	0	0	0	0	8	216	269	161	99
3.25	0	0	0	0	0	37	165	182	205
3.5	0	0	0	0	2	40	181	223	285
3.75	0	0	0	0	1	12	24	34	84
4.0	0	0	0	0	0	12	18	22	84
4.25	0	0	0	0	0	3	7	9	60
4.5	0	0	0	0	0	8	14	15	63
4.75	0	0	0	0	0	1	2	1	19
5.0	0	0	0	0	0	3	2	0	16
5.25	0	0	0	0	0	2	0	2	9
5.5	0	0	0	0	0	0	0	0	10
5.75	0	0	0	0	0	0	1	0	3
6.0	0	0	0	0	0	0	1	0	5
6.25	0	0	0	0	0	0	0	0	2

- There is a clear change when grade goes up to 8. Grade less than 8 is more likely to have 1 floor while grade greater than 7 is more likely to have 2 floors.

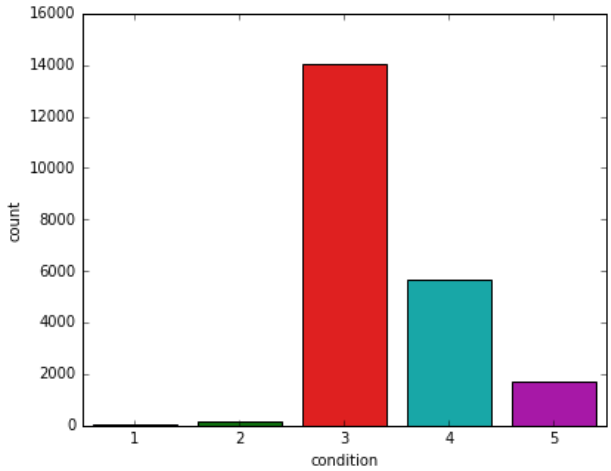
grade	1	3	4	5	6	7	8	9	10
floors									
1.0	1	3	27	202	1662	5916	2233	447	189
1.5	0	0	2	38	311	1006	402	105	46
2.0	0	0	0	2	63	1943	2989	1935	1309
2.5	0	0	0	0	2	15	53	46	45
3.0	0	0	0	0	0	100	385	82	46
3.5	0	0	0	0	0	1	6	0	1

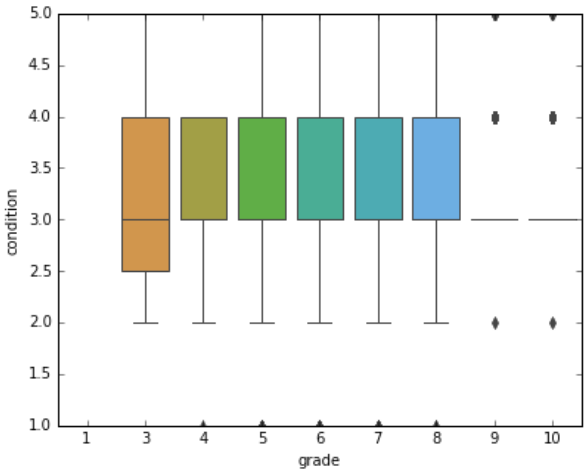
- more houses built in 2012 than other years. Houses with grade less than 6 are more likely to be houses built before 1980.





To proof the weak relationship between grade and those unrelated features found in heat map, I will take the condition as an example and will check its distribution. Condition has the same distribution for grade score from 3 to 8. It indeed has not strong relationship with grade.



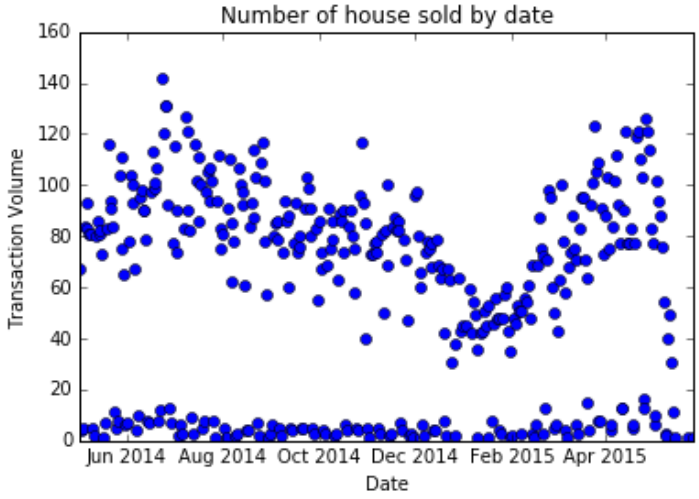


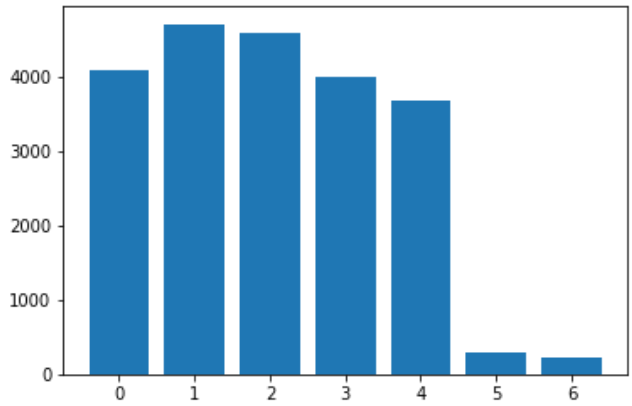
What are the busiest times of the year to buy/sell house in Seattle?

From the number of houses sold by date, we can see, In General, the transaction volume dropped to the lowest point in winter and reached the highest point in summer.

There are some low transaction days evenly distributed along the year. By checking the transaction volume by day of the week,

we can the low points are the weekend. And interestingly, Tuesday is the day with most transactions.





Predict house grade based on the given features

I Prepare training and test data before we jump into algorithms. From the section 1, I learnt that the top 4 features having the strongest relationship with grade are sqft_above, sqft_living15, bathrooms, price. So I will test the full set of feature with only the four features to train the model to compare the efficiency.

I split the data into 80% training and 20% testing data set. I used standard scaler on the feature columns, it increased the score from 0.55 to 0.65.

As proposed, I started with linear regression model, and compare it with more advanced and embedded methods including KNN, Adaboost classifier, Decision tree regressor, SVM, Random Forest Regressor and Extreme Gradient Boosting Regressor for comparison.

Method	R2 Score
Linear Regression	0.69
KNN	0.65
AdaBoost Classifier	0.47
Decision Tree	0.63
SVM	0.77
Random Forest Regressor	0.79

I also tested with Extreme Gradient Boosting.

Grade	precision	recall	f1-score	support
3	0.00	0.00	0.00	1
4	0.33	0.20	0.25	5
5	0.60	0.11	0.18	56
6	0.67	0.50	0.57	426
7	0.74	0.82	0.78	1787
8	0.64	0.64	0.64	1191
9	0.54	0.52	0.53	513
10	0.73	0.70	0.71	344
avg / total	0.68	0.69	0.68	4323

Interesting to see linear algorithm performs better than KNN in respect of the R2 score. KNN is better than ada boost classifier. Among regressor algorithms, random forest regressor has the highest r2 score.

By cutting the features down to the top 7 key features, we can see the linear model score dropped from 0.69 to 0.68 which is acceptable. While for the Random forest regressor, the R2 score dropped from 0.79 to 0.75 by cutting off features down.

I will use the random grid to find the best hyperparameters for random forest regressor to increase the score for random forest regressor.

Refinement

The searching grid find the best hyperparameter as follows 'n_estimators': 800, 'min_samples_split': 4, 'min_samples_leaf': 1, 'max_features': 'log2', 'max_depth': 30.

It increased the score to 0.92 which is very optimistic.

Conclusion

here is what I have learned from this project.

1. sqft_above, sqft_living15, bathrooms, price are the four key features determine the house grade.
2. Seattle people like to buy/sell house in summer. And by week, Tuesday is the day with most transactions.
3. Random forest regressor model has the highest R2 score comparing to linear regression, SVM, KNN, XGBoost and adabooster algorithms. I used grid searching and it pumped score from 0.7 to 0.9.

Reflection

The major process steps are

1. Find the initial public relevant dataset
2. explored and preprocessed the data (lots of learning in visualizations)
3. created benchmark classifier
4. trained the classifier using the training data. (most interesting section with lots of try out)
5. pick up the best performer algorithm based on the R2 score.
6. searching grid to find the best hyperparameter with random forest regressor.
7. test out the model with searching grid result.

Something interesting to find is linear regression worked better than ada bosster with default parameter.

By adding more features will not improve the model a lot if the features are redundant. e.g. sqrt_living and sqrt.

calculation load is not a big problem here given only 21.6k records. except SVM, I did not see a big difference in calculation time. The searching grid takes some time to try out the best combination.

next step

I would like to test the model with some more data in the future if I can find similar data after 2015.

Reference

[1]: https://blog.exploratory.io/introduction-to-extreme-gradient-boosting-in-exploratory-7bbec554ac7

[2]: https://simple.wikipedia.org/wiki/Random_forest