Machine Learning for House Price Prediction: A Multiple Model Approach

Satyawant Kumar – 2017118062

Abstract--- **Machine learning plays a major role from past years in image detection, spam reorganization, normal speech command, product recommendation and medical diagnosis. Present machine learning algorithm helps us in enhancing security alerts, ensuring public safety and improve medical enhancements. Machine learning system also provides better customer service and safer automobile systems. In the present paper we discuss about the prediction of future housing prices that is generated by machine learning algorithm. Usually, House price index represents the summarized price changes of residential housing. While for a single family house price prediction, it needs more accurate method based on location, house type, size, build year, local amenities, and some other factors which could affect house demand and supply. For the selection of prediction methods we compare and explore various prediction models. We utilize "Random-Forest" as our final deployment model. Our result exhibit that our approach of the issue need to be successful, and has the ability to process predictions that would be comparative with other house cost prediction models. More over on other hand housing value indices, the advancement of a housing cost prediction that tend to the advancement of real estate policies schemes. We utilizes machine learning algorithms that develops housing price prediction models. We in that point recommend a housing cost prediction model to support a house vender or a real estate agent for better information based on the valuation of house. Examinations exhibit that "Random-Forest" algorithm, in view of accuracy, reliably performs better than alternate models in the execution of housing cost prediction.**

Index Terms--- **Machine learning algorithm, Random-Forest-Regressor, Decision-Tree-Regressor and Linear Regression, hosing cost prediction, R-Squared Error, Cross Validation.**

1. INTRODUCTION

Development of civilization is the foundation of increase of demand of houses day by day. Accurate prediction of house prices has been always a fascination for the buyers, sellers and for the bankers also. Many researchers have already worked to unravel the mysteries of the prediction of the house prices. There are many theories that have been given birth as a consequence of the research work contributed by the various researchers all over the world. Some of these theories believe that the geographical location and culture of a particular area determine how the home prices will increase or decrease whereas there are other schools of thought who emphasize the socio-economic conditions that largely play behind these house price rises. We all know that house price is a number from some defined assortment, so obviously prediction of prices of houses is a regression task. To forecast house price one person usually tries to locate similar properties at his or her neighbourhood and based on collected data that person will try to predict the house price. All these indicate that house price prediction is an emerging study area of regression which requires the knowledge of machine learning. This has motivated to work in this domain.

Machine learning develops algorithms and builds models from data, and uses them to predict on new data. The main difference with traditional algorithm is that a model is built from inputs data rather than just execute a series of instructions. Supervised learning uses data with result labeled, while unsupervised learning using unlabeled data. There are a few common machine learning algorithms, such as regression, classification, neural network and deep learning. Reinforcement learning and representation learning are heavily used for deep learning.

How to use machine learning algorithms to predict house price? It is a challenge to get as closely as possible result based on the model built. For a specific house price it is determined by location, size, house type, city, country, tax rules, economic cycle, population movement, interest rate, and many other factors which could affect demand and supply. After examining data, we find that the data quality is a key factor to predict the house prices. Data input feature density estimation is important for regression.

This paper is organized as follows, it reviews related previous work in part II, data visualization & splitting in part III, and illustrates the details of different models designed in part IV, then compares the test result of different algorithms in part V, finally discusses the result and makes a conclusion in part VI.

2. RELATED WORKS

A few other documents explore the correlation among house price and local amenities, local area and renovation. Eli Beracha [1] investigate the correlation between house price volatility, returns and local amenities, and proves that high amenity areas experience greater price volatility. Stephen Law [2] finds that the strong links between Street-based local area with house price and it shows that using Street-based local is better than using region-based local area. Alexander and William investigates the result of property improvements in wide-scale US geographies [3], the result shows that the price could be increased 15% in the central districts of large cities, while less distortionary effect outside of downtown areas or in smaller cities.

PROPOSED SYSTEM

Nowadays, e-education and e-learning is highly influenced. Everything is shifting from manual to automated systems. The objective of this project is to predict the house prices so as to minimize the problems faced by the customer. The present method is that the customer approaches a real estate agent to manage his/her investments and suggest suitable estates for his investments. But this method is risky as the agent might predict wrong estates and thus leading to loss of the customer's investments. The manual method which is currently used in the market is out dated and has high risk. So as to overcome this fault, there is a need for an updated and automated system. Data mining or machine learning algorithms can be used to help investors to invest in an appropriate estate according to their mentioned requirements. Also the new system will be cost and time efficient. This will have simple operations. The proposed system works on Random Forest Algorithm.

3. EXPLANATERY DATA ANALYSIS

3.1 Description Of Data

1.	**id**	Unique ID for each home sold
2.	**date**	Date of the home sale
3.	**Price**	Price of each home sold
4.	**bedrooms**	Number of bedrooms
5.	**bathrooms**	Number of Bathrooms, where 0.5 accounts for a room with a toilet but no shower
6.	**Sqft_living**	Square footage of the apartments interior living space
7.	**Sqft_lot**	Square footage of the land space
8.	**floors**	Number of floors
9.	**waterfront**	Dummy variable for whether the apartment was overlooking the waterfront or not
10.	**view**	An index from 0 to 4 of how good the view of the property was
11.	**condition**	An index from 1 to 5 on the condition of the apartment
12.	**grade**	An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality of construction and design.
13.	**Sqft_above**	The square footage of the interior housing space that is above ground level
14.	**Sqft_basement**	The square footage of the interior housing space that is below

		ground level
15.	**Yr_built**	The year the house was initially built
16.	**Yr_renovated**	The year of the house’s last renovation
17.	**zipcode**	What zipcode area the house is in
18.	**lat**	Lattitude
19.	**long**	Longitude
20.	**Sqft_living15**	The square footage of interior housing living space for the nearest 15 neighbors
21.	**Sqft_lot15**	The square footage of the land lots of the nearest 15 neighbors

This dataset consists of 19 house features (excluding “ID” & “date”) and 21613 houses with sold prices. Although the dataset is relatively small with only 21613 examples, it contains 19 features such as areas of the houses, number of the floors, and numbers of bathrooms etc. Such large amounts of features enable us to explore various techniques to predict the house prices.

The dataset consists of features in various formats. It has numerical data such as prices and numbers of bathrooms/bedrooms as well as categorical features such as waterfront which can be seen from the above table.

3.2 Relationship between variables

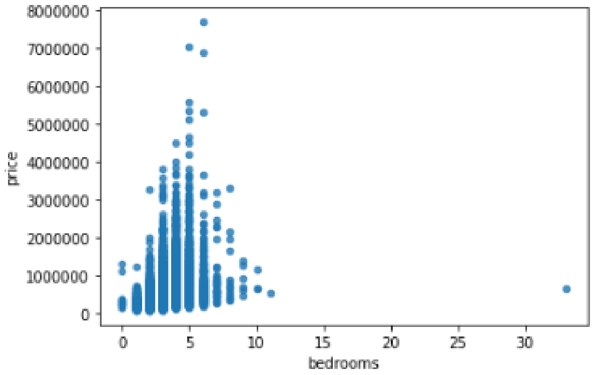


Fig 1: bedrooms V/s price

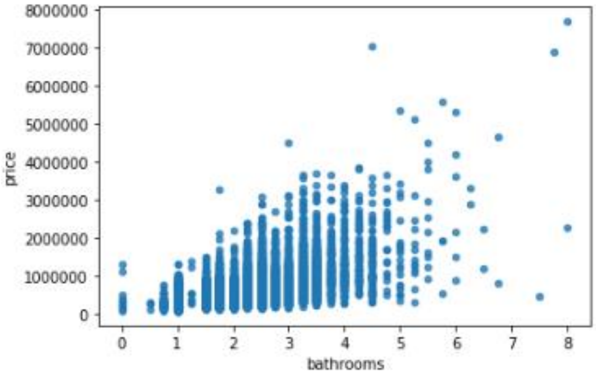


Fig 2: bathrooms V/s price

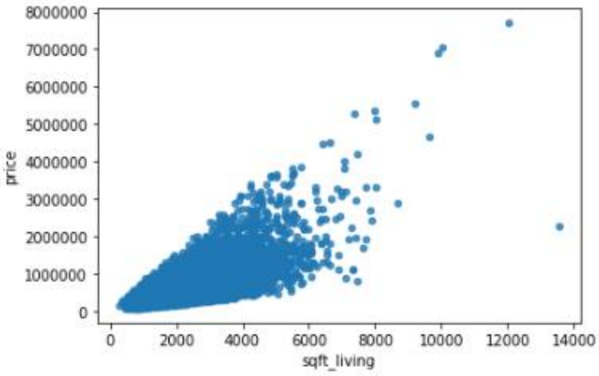


Fig 3: Sqft_living V/s price

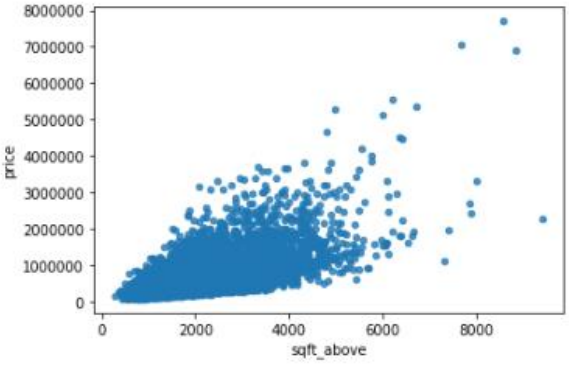


Fig 4: Sqft_above V/s price

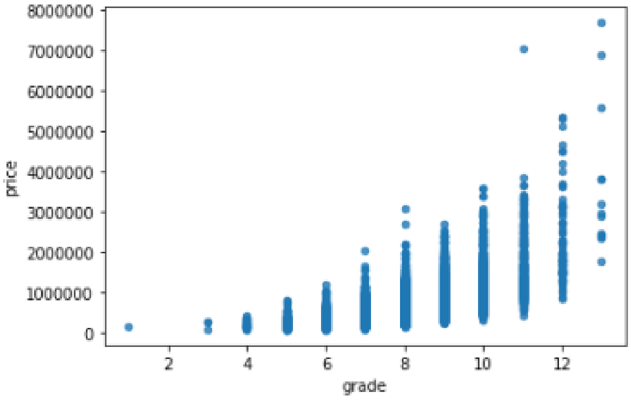


Fig 5: grade V/s price

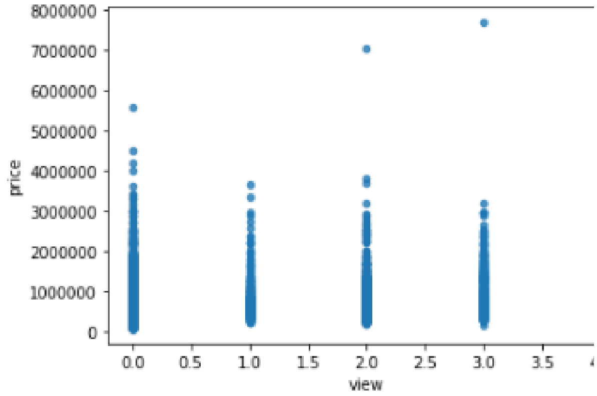


Fig 6: View V/s price

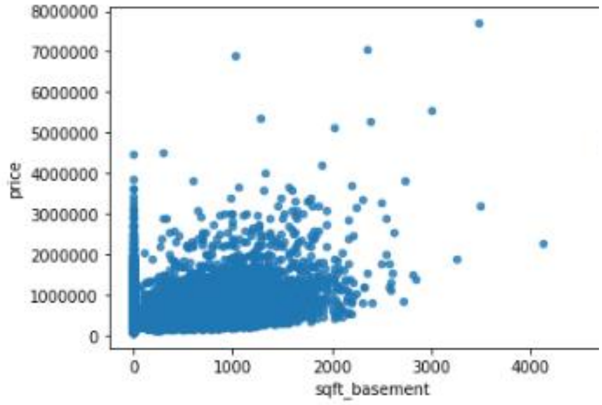
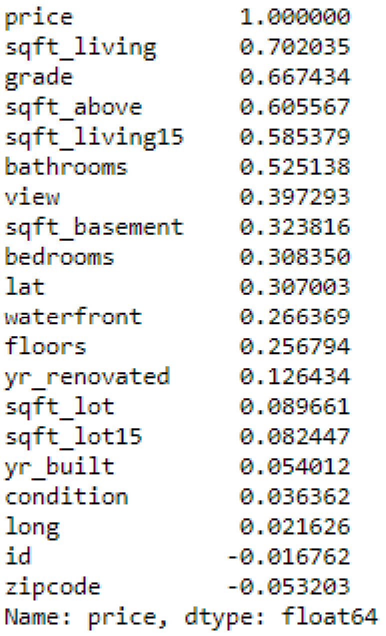


Fig 7: Sqft_basement V/s price



```
price            1.000000
sqft_living      0.702035
grade            0.667434
sqft_above       0.605567
sqft_living15    0.585379
bathrooms        0.525138
view             0.397293
sqft_basement    0.323816
bedrooms         0.308350
lat              0.307003
waterfront       0.266369
floors           0.256794
yr_renovated     0.126434
sqft_lot         0.089661
sqft_lot15       0.082447
yr_built         0.054012
condition        0.036362
long             0.021626
id              -0.016762
zipcode         -0.053203
Name: price, dtype: float64
```

Fig 8: Co-relation of price with other parameters

From the above Co-relation figure 8 we can depict that parameters **like "sqft_living", "grade", "sqft_above", "sqft_living15", "bathrooms"** are highly co-related with the target parameter **"price".**

In order to make this data with different format usable for our algorithms, we have excluded some features from the dataset which are least co-related with the target variable "Price". The final dataset has 13 features, in which 12 are the predictor variables and one "price" is the target variable.

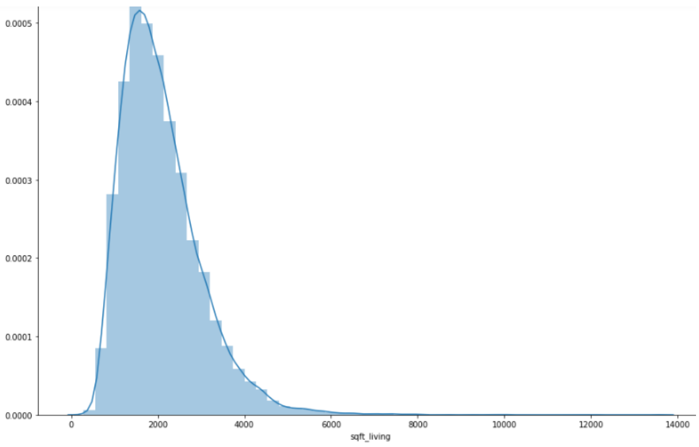


Fig 9: distplot for sqft_living

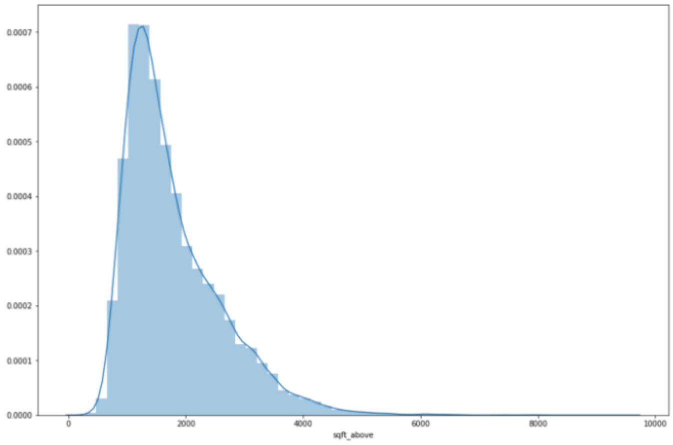


Fig 10: distplot for sqft_above

3.3 Training and Testing dataset

What is the benefit to splitting a dataset into some ratio of training and testing subsets for a learning algorithm?

It is useful to evaluate our model once it is trained. We want to know if it has learned properly from a training split of the data. There can be 3 different situations:

1) The model didn´t learn well on the data, and can't predict even the outcomes of the training set, this is called underfitting and it is caused because a high bias.

2) The model learn too well the training data, up to the point that it memorized it and is not able to generalize on new data, this is called overfitting, it is caused because high variance.

3) The model just had the right balance between bias and variance, it learned well and is able predict correctly the outcomes on new data.

We divided our dataset into training and testing set with a roughly 80/20 ratio, with 17290 training examples and 4323 testing examples.

We have used a "SimpleImputer" from sklearn.impute to fill the empty values with the median.

4. MODELS

We have created three different models using machine learn algorithms: **Regression, Decision Tree and Random Forest**. While it seems more reasonable to perform regression since house prices are continuous in nature. Classifying house prices into individual ranges of prices would also provide helpful insight for the users; also, this helps us to explore different techniques which might be regression or classification-specific even.

4.1 Linear Regression

There are many regression algorithms that can be used to build models and predict house prices. Here, we have used simply "Linear Regression".

For this model, we try to solve the following problem: given a processed list of features for a house, we would like to predict its potential sale price. Linear regression is a natural choice of baseline model for regression problems. So we first ran linear regression including all features, using our 12 features and 17290 training samples. The model is then used to predict sale prices of houses given features in our test data and is compared to the actual sale prices of houses given in test data set.

4.1.1 Model Evaluation Results

The model has generated **Sqaured mean erro r 216731.38, R squared training 0.662 and R sqaured testing 0.658.**

Cross Validation Method – Using this meth od, linear model has generated a **Mean of 10 RMSE = 212640.367 and Standard deviatio n = 18909.5614.**

4.2 Decision Tree

A tree has many analogies in real life, and turns out that it has influenced a wide area of **machine learni ng**, covering both **classification and regression**. In decision analysis, a decision tree can be used to visu ally and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions. Though a commonly used tool in data mining for deriving a strategy to reach a particular g oal, it is also widely used in machine learning. After linear regression we created another model ba sed on decision tree. Like in the previous model, her e also we have used 12 features or predictor variable s and 17290 training samples. The model is then use d to predict sale prices of houses given features in o ur test data and is compared to the actual sale prices of houses given in test data set.

4.2.1 Model Evaluation Results

The model has generated **Sqaured mean erro r 229015.64, R squared training 0.999 and R sqaured testing 0.618.**

Cross Validation Method – Using this meth od, linear model has generated a **Mean of 10 RMSE = 216495.973 and Standard deviatio n = 8158.403.**

4.3 Random Forest

Random forest, like its name implies, consists of a l arge number of individual decision trees that operat e as an ensemble. Each individual tree in the rando m forest spits out a class prediction and the class wi th the most votes becomes our model's prediction (see figure below).

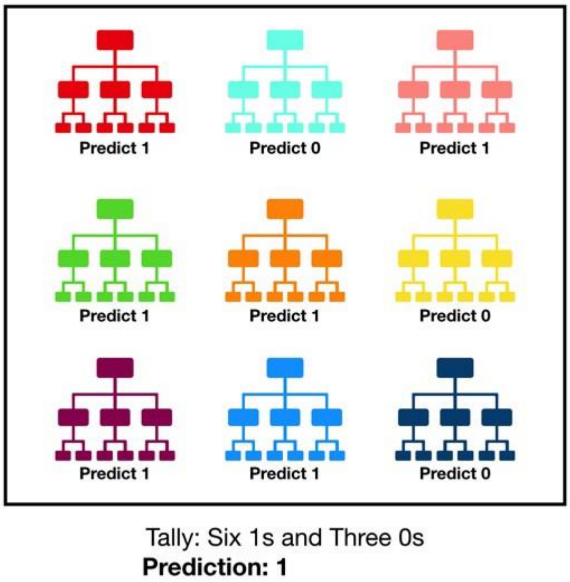


Fig 11: Random Forest Model

4.3.1 Model Evaluation Results

The model has generated **Sqaured mean error 169692.49, R squared training 0.974 and R sqaured testing 0.79.**

Cross Validation Method – Using this method, linear model has generated a **Mean of 10 RMSE = 155150.539 and Standard deviation = 11359.642.**

4.4 Performance Measurement Methods

It is difficult to measure the quality of a given model without quantifying its performance on the training and testing. This is typically done using some type of performance metric, whether it is through calculating some type of error, the goodness of fit, or some other useful measurement. Below are some methods which we have used to measure the performance of our models.

4.4.1 R Squared

For this project, we will calculate the coefficient of determination, R², to quantify the model's performance. The coefficient of determination for a model is a useful statistic in regression analysis, as it often describes how "good" that model is at making predictions.

The values for R² range from 0 to 1, which captures the percentage of squared correlation between the predicted and actual values of the target variable.

4.4.2 RMSE (Root Mean Squared Error)

Root mean squared error (RMSE) is the square root of the mean of the square of all of the error. The use of RMSE is very common, and it is considered an excellent general-purpose error metric for numerical predictions.

4.4.3 Cross Validation

K-fold cross-validation is a technique used for making sure that our model is well trained, without using the test set. It consist in splitting data into k partitions of equal size. For each partition i, we train the model on the remaining k-1 parameters and evaluate it on partition i. The final score is the average of the K scores obtained.

5. MODEL COMPARISION

Model	Sqaured mean error	R squared training	R sqaured testing
Linear Regression	216731.38	0.662 = 66.2%	0.658 = 65.8%
Decision Tree	229015.64	0.999 = 99.9%	0.618 = 61.8%
Random Forest	169692.49	0.974 = 97.4%	0.79 = 79 %

Table 1: Squared mean error & R-Squared

Model	RMSE	Std.Deviation
Linear Regression	212640.367	18909.5614
Decision Tree	216495.973	8158.403
Random Forest	155150.539	11359.642

Table 2: Cross Validation Results

5.1 Selection Of Best Model

So, from the above result we can depict that R-squared testing of random forest model is 79% which is greater as compared to both the other models. Also, through cross validation method we got least RMSE of 155150.539 against other two models. **In this way we can say that "Random Forest" performs better than other models**.

5.2 Making Prediction

Once a model has been trained on a given set of data, it can now be used to make predictions on new sets of input data**.** Here, we have predicted price of house based on 5 or 6 different set of features taken from test datasets.

5.3 Model's Sensitivity

An optimal model is not necessarily a robust model. Sometimes, a model is either too complex or too simple to sufficiently generalize to new data.

Sometimes, a model could use a learning algorithm that is not appropriate for the structure of the data given.

Other times, the data itself could be too noisy or contain too few samples to allow a model to adequately capture the target variable — i.e., the model is underfited.

5.4 Model's Applicability

Now, we use these results to discuss whether the constructed model should or should not be used in a real-world setting. Some questions that are worth to answer are:

How relevant today is data that was collected from 1978? How important is inflation?

Data collected from 1978 is not of much value in today's world. Society and economics have changed so much and inflation has made a great impact on the prices.

Are the features present in the data sufficient to describe a home? Do you think factors like quality of appliances in the home, square feet of the plot area, presence of pool or not etc should factor in?

The dataset considered is quite limited, there are a lot of features, like the size of the house in square feet, the presence of pool or not, and others, that are very relevant when considering a house price.

Is the model robust enough to make consistent predictions?

Depending upon the variance of the price range, we can assure that it is robust or not a robust model and, therefore, appropriate for making predictions or not.

Would data collected in an urban city like Boston be applicable in a rural city?

Data collected from a big urban city like Boston would not be applicable in a rural city, as for equal value of features prices will be much higher in the urban area.

Is it fair to judge the price of an individual home based on the characteristics of the entire neighbourhood?

In general it is not fair to estimate or predict the price of an individual home based on the features of the entire neighbourhood. In the same neighbourhood there can be huge differences in prices.

6. Results & Conclusion

We made a machine learning regression project from end-to-end and we learned and obtained several insights about regression models.

In today's real estate world, it has become tough to store such huge data and extract them for one's own requirement. Also, the extracted data should be useful. The system makes optimal use of the Random Forest Algorithm. The system makes use of such data in the most efficient way. The algorithm helps to fulfil customers by increasing the accuracy of estate choice and reducing the risk of investing in an estate. A lot's of features that could be added to make the system more widely acceptable. One of the major future scopes is adding estate database of more cities and features which will provide the user to explore more estates and reach an accurate decision. More factors like recession that affect the house prices shall be added. In-depth details of every property will be added to provide ample details of a desired estate. This will help the system to run on a larger level.

REFERENCES

[1] Eli Beracha, Ben T Gilbert, Tyler Kjorstad, Kiplan womack, "On the Relation between Local Amenities and House Price Dynamics", Journal of Real estate Economics, Aug. 2016.

[2] Stephen Law, "Defining Street-based Local Area and measuring its effect on house price using a hedonic price approach: The case study of

Metropolitan London", Cities, vol. 60, Part A, pp. 166–179, Feb. 2017.

[3] Alexander N. Bogin , William M. Doerner, "Property Renovations and Their Impact on House Price Index Construction", https://www.fhfa.gov/PolicyProgramsResearch/Research/PaperDocuments/wp1702.pdf