Introduction to Data Science Tools & Techniques

## Group Assignment No. 5

# House Price Prediction

## 

### Group Members:

Ali Nauman – 18L1863  
Hammad Akram – 18L1808  
Ammar Hussain – 18L1834

## Problem Statement

Ask a home buyer to describe their dream house, and they probably won’t begin with the height of the basement ceiling or the proximity to an east-west railroad. But this dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

The main idea behind this problem is to predict the price of any particular house based on the features that are included in the data-set. There are multiple regression techniques that can be utilized to predict accurate prices based on the features extracted and manipulated.

Also, the model’s accuracy is to be determined by calculating the **Root Mean Square Error** (RMSE).

### Root Mean Square Error

Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.

The formula is:



**Where**:

* f = forecasts (expected values or unknown results),
* o = observed values (known results).

In data science, RMSE serves as a double purpose:

1. To serve as a heuristic for training models
2. To evaluate trained models for usefulness and accuracy

## Model Implementation

We will be utilizing Linear Regression and Random Forest Regression as the two models to predict ‘*price*’ of the house based on the features we select.

### Linear Regression

Linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models.

### Random Forest Regression

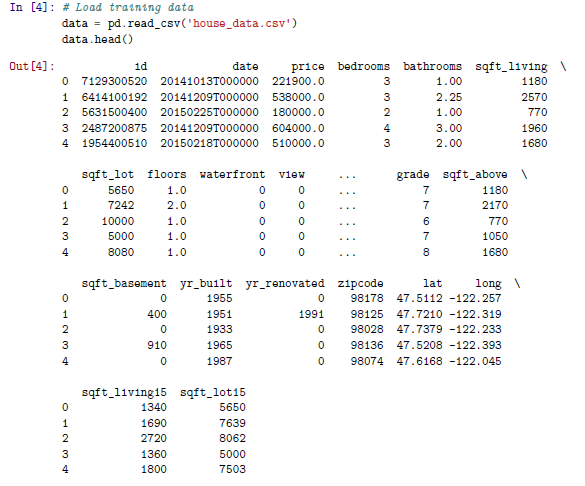
Random forest is a Supervised Learning algorithm which uses ensemble learning method for classification and regression.

Random forest is a bagging technique and not a boosting technique. The trees in random forests are run in parallel. There is no interaction between these trees while building the trees. It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

## Experiment Set-Up

### Exploring the Data

The first thing you need to do before solving any Data Science problem is getting familiar with the data-set. Get to know your data by printing out some stats, checking its dimensions and checking data types of features.



The data-set has 21 variables including an ‘*ID*’ column and a ‘*date*’ column. The shape of the data-set is (21613, 21).

### Data Wrangling & Preprocessing

We have to preprocess our data in order to make it useful for data analysis and model training. Although, the steps involved vary depending on the problem and the dataset but here we have followed a roughly generic approach which is applicable for most problems. The steps involved are as follows:

#### Look for Null or Missing Values

The data-set upon investigation does not contain any null or missing values

#### Change data types of features

All the features listed in the data-set are of the type, *int*. Only, ‘*date*’ is being provided as a categorical data type, which has been converted into date-time.

#### Encode data of categorical features

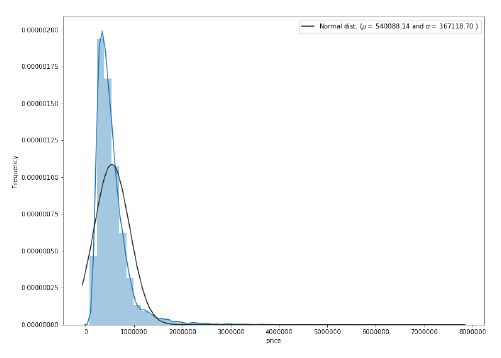
As we do not have any important feature, which was of the type, categorical, so we this step was not required.

### Data Analysis and Visualizations

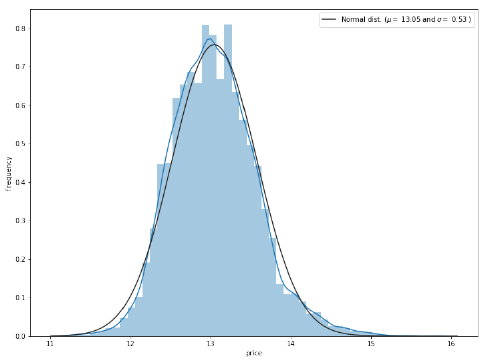
Performing a detailed analysis of the data helps you understand which features are important, what’s their correlation with each other which features would contribute in predicting the target variable. Different types of visualizations and plots can help you achieve that. The ’Target Variable’ in this data is the price column.

We will now perform some analysis on the target variable to get a better insight into what we are working on.

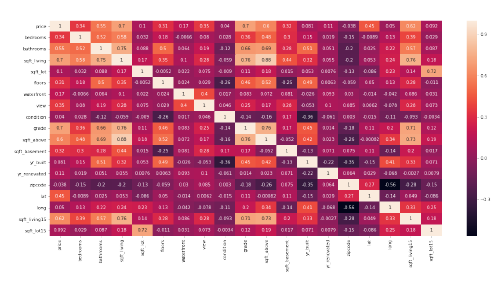
Starting off, we will evaluate the distribution the price of houses as following.



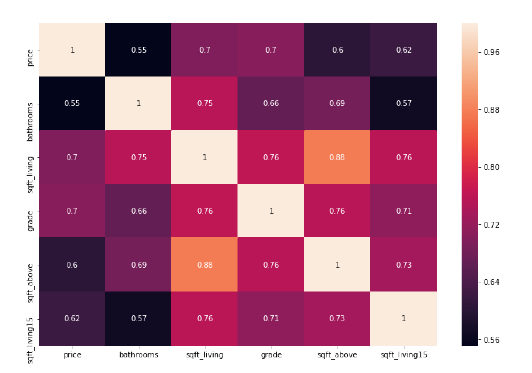
This target variable is right skewed. Now, we need to transform this variable and make it as a normal distribution. Here we will use *log* for target variable to make it a normal distribution.

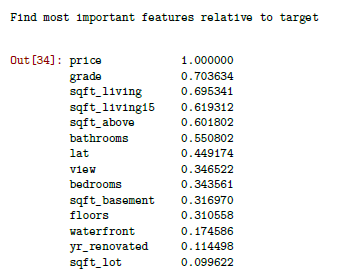


Moving forward, we need to find the correlation between the attributes. This will provide us with a better understanding of all the attributes and how they are in relevance to predicting the price of the house.



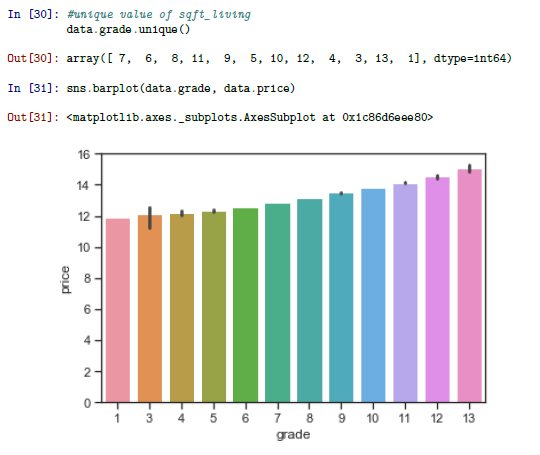
Also, we have further evaluated top 50% correlation of the attributes with the price,



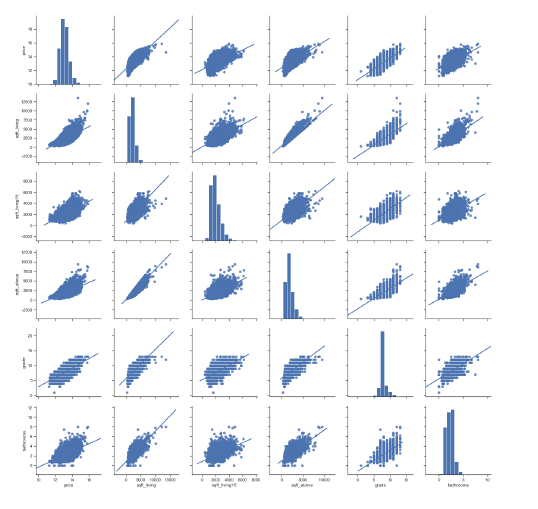


From this analysis we can safely evaluate that the attribute ‘*grade*’, is highly correlated with the target feature of price by 71%.

We further evaluated the unique values in the attribute ‘*grade*’ and plotted a bar plot to check the significance as well.



From the earlier investigation of the top features that are contributing the most, we displayed there relevance to each other and the target feature, ‘*price*’ and how it is affecting it.



As a final step, we imputed the attribute ‘*date*’, as it showed no relevance to the data-set and in no way was it affecting the target feature, ‘*price*’.

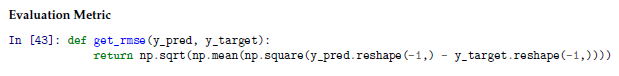
## Model Training

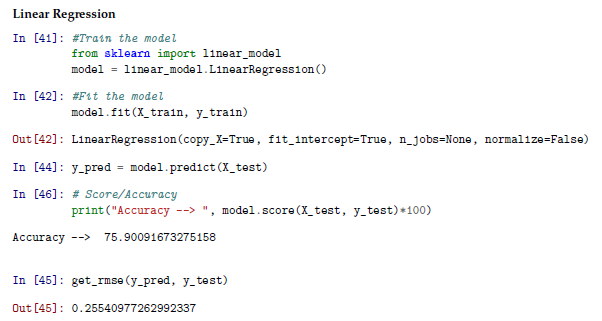
This is a regression problem since we are predicting house prices which is a continuous random variable. The steps involved are as follows:

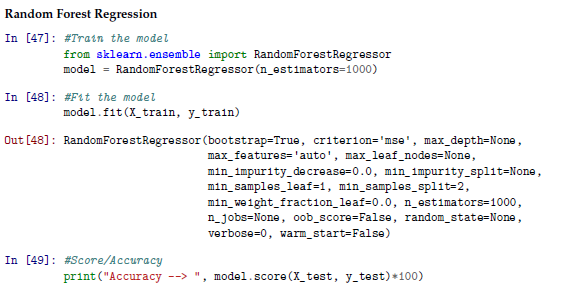
* Standardize or Normalize Training Data
* Train Test Split
* Train Model
* Evaluation based on RMSE

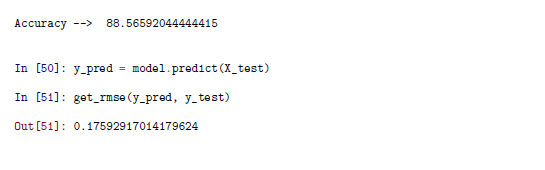
We have implemented both the *Linear Regression* and *Random Forest Regression* model to predict the values and evaluate the performance based on the RMSE scores.

**Root Mean Square Error** is defined as following,









## Results

Please find below the table which include the results including the **RMSE** and **Model Score** for both regression models,

|  |  |  |
| --- | --- | --- |
| **Model** | **RMSE** | **Model Score** |
| **Linear Regression** | 0.255 | 75.9% |
| **Random Forest Regression** | 0.176 | 88.6% |

*Random Forest Regression* has an **RMSE** value of 0.176 as compared to *Linear Regression*, 0.255.

## Conclusion

From the results computed from both the models, we can see that the **RMSE** and **Model Score** of *Random Forest Regression* is much better than the *Linear Regression*. This is mainly due to the reason that the *Random Forest* utilizes *Ensemble Learning* technique, which is a boosting and bagging phenomena, to perform better evaluations on the given data.