Linear Regression Model

How linear regression works.

- Linear regression model predicts the value of a dependent variable using one or more predictor variables.
- Before applying the model, we need to first determine the correlation between variables. If the variables have no correlation, the Regression Model will fail and there is no advantage of including this predictor variable in the regression model. [1]
- A linear equation of the form Y = mX + c is defined in between the two variables, where Y = dependent variable and X= list of predictor variables. The line is plotted on the observed data.
- This is a deterministic relationship. But in real life, we do not find such a relationship between two variables. Instead we are always looking for Statistical relationships where the relationship between variables is not perfect. There are always points on the plot below or above our line plot. There can be outliers too.[2]
- R-squared value is measure of the model for variables X and Y. Higher the r-squared value means our model is able to explain high number of data, i.e. more data points (Y,X) lie on the line plotted by our model. While a zero or low value of r-squared means that predictor variables in X does not influence the dependent variable Y.
- Mean Squared Error (MSE) shows the difference between the actual value and the predicted value. It is the average of the squares of the deviations. Lower MSE score is better.
- The linear regression plot by statsmodels ols api for pandas provides a summary which includes the above discussed measures of model efficacy.

References:

1. Linear Regression [http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm]
2. Simple Linear Regression [https://onlinecourses.science.psu.edu/stat501/node/251]

An evaluation of how well Linear Regression model works.

- The R-squared value is 0.008 without intercept, which means that only 0.8% of the log errors can be explained by our predictor variables bedroomcnt, calculatedfinishedsquarefeet and garagetotalsqft.
- As the dependent variable logerror is a very small value for each result, coefficient values for each predictor variable (provided by statsmodels ols api) are also very small. This means that even a major increase in the value of our predictor variable will only cause a small change to logerror.
- Though for every predictor variable the standard error is very low. But to confidently say this, metric evaluation of the variables needs to be performed.
- Even after taking the z_score of the predictor variables, the results remain the same. So, normalization does not affect our model.
- The r-squared value is low because even small improvements in the zillow model are hard to come by as it is already a very good model
- The Linear Regression model turns out to be better than K-Nearest Neighbor Model (MSE of 0.02) and Decision Tree Regressor Model (MSE of 0.05)
- Kaggle submission with linear regression model gives a score of 0.0649077 and rank of 2041

Any interesting experiences or surprises you had over the course of these experiments.

- Data Cleaning is definitely one of the important aspects of designing a model. Without properly cleaned dataset, no matter how good a model is, it cannot perform well. Replacing NaN values, breaking categorical values to binary columns, removing redundant columns are some of the data cleaning techniques we can apply to a database before we start modeling.
- We need to KISS (Keep it simple, stupid). Starting off with a very basic linear regression model is always a good idea before moving on to a more complex model. We need to understand where the basic model lacks, shortlist the points we need to improve on and then make a better model.
- Choices of predictor variables play a huge role in any model. Choosing variables with good correlation will help make a better model while adding variables with

very less correlation will only add redundant data to the model. So, a correlation analysis is a must. Graph plots help in visualizing the correlated data pictorially. Different plots serve specific unique purpose, like scatter plots can help in determining outliers and thus correcting errors in our dataset.

- There is always a need for a testing dataset so as to check the efficacy of our model. So, it is good practice to break-up the original dataset into a training data and a testing data.
- The complex model will not necessarily give better results. As in our case, linear regression model has better MSE values than both K-Nearest neighbor and Decision tree model. The scores received from Kaggle also imply the same.