Homework 3: Data Integration and Modelling

In predicting the log error two models were used.

- Linear Regression
- Random Forests

Among these, Linear Regression performed well in processing and predicting the log errors of the data given.

Linear Regression:

A linear approach in predicting a value as a function of n-variables. Can be written as:
y = a+bx ; where y is the dependent variable and x being the independent variable.

If x is scalar i.e, one variable, the regression is known as simple linear regression.
If x is a vector i.e, more than one variable, the regression is known as multiple linear regression.

Once the model is trained with the observed set of y values for given x values, it will be able to predict the y value for a new x value.

The ***sklearn*** library of python has inbuilt functions to apply the linear regression on a set of data and predict the results. The function score return the co.efficient of determination R^2 of the prediction. The value varies from +ve to -ve. +ve value indicates the rise towards the better fit of the model (the best value being 1). Negative value indicates that the model is not a better fit.

This model has been trained and fit with 16 parameters to our data to predict the log error. Based on the Pearson Correlation Coefficient from homework 2, the positively corelated variables were selected in the initial step. These parameters include the external data i.e, the crime data (in our case) which was not provided by the zillow web site.

Random Forests:

Random Forests works by constructing decision trees during training and out putting the mode of mean prediction of individual trees.

Results from Kaggle:

1. Linear Regression

----Best score of linear regression---- Mean Squared Error : 0.01984 score by kaggle : 0.065 Rank : 2758

The mean squared error from home work 2 where we didnot use any external data was : 0.028 This error was reduced further after adding the external data i.e, crime data. The new mean

squared error : 0.01984. We can see that the model performed better than the previous iteration. But the kaggle rank did not improve much even though the mean squared error was reduce.

2. Random Forests

----Best score of random forest regressor---- Mean Squared Error : 0.0229 Score by kaggle : 0.088

The random forest regression did not perform better than linear regression.

We also tried using k nearest neighbours . But this failed because of memory error.

Out of all the iterations with different models, we found that linear regression performed well with a minimum mean squared error. Also the time taken to predict the log error was comparatively very less than the other model. This indicates the better performance of linear regression with the size of data.