Project Report
Homework 2: Exploratory Data Analysis in IPython

This report discusses the different models used to solve the Zillows Challenge for predicting the log error of different properties on sale.

The data set consisted of the Properties file describing all the various properties about the parcel on sale which Zillow uses to predict the sales price. This data consisted of categorical,discrete and continuous values.

Following are the models used:

1. Linear Regression Model
   Linear Regression follows a linear approach in modelling the relationship between scalar dependent variable 'Y' and one or more independent variables denoted by 'X'.

   $$Y = mX + e$$

   Where 'm' is the linear coefficient array which is multiplied to the X data set.
   And 'e' is the noise or induced error value.

   Linear relationships are the simplest relationship that can be imagines between any 2 variables. Their simplicity makes it widely used across multiple applications in Statistics, Biology etc. Even if the variables are not linear, we can apply some transformation and reduce the gap between them.

2. Random Forest Regression Model
   Random Forest model is part of ensemble learning methods which works by constructing multiple decision trees and outputting the class which is the mode or mean prediction of the decision tree.

Analysis

In the Zillows problem surprisingly Linear Regression and Random Forest both gave same mean square error, however compared to the coefficients in Linear the Feature importance values in Regression were a better estimation on the relation of these features with the log error.

Some of the attributes were highlighted in both the models as the most important ones like totalarea, taxamount , bedrooomcnt. Thus concluding that these variables play a major role in determining the log error values.

Also one more interesting relationship that Random Forest model predicted was between latitude and log error. Thus suggesting that a parcel's location also plays a vital role in the selling price.

Also the Random Forest Regression model performed better when the features in the training data were reduced after the results of the previous regression values where most of the features were part of the training model.

Hence I feel Random Forest Regression model with the correct and appropriate features and their values is a good model to predict as the decision trees can be pruned with the desired depth, weight and number of leaves. Thus giving more options to fine tune the results.