CSE 519 **Neha Mane**
HW2 - Report Date: 09/27/2017

HW2 Statement

Exploratory Data Analysis in IPython based on Zillow Prize Challenge on Kaggle

Regression Models

I have used Linear Regression for question 3 and Random Forest Regression as the advanced model for creating the final predicted csv

Linear Regression

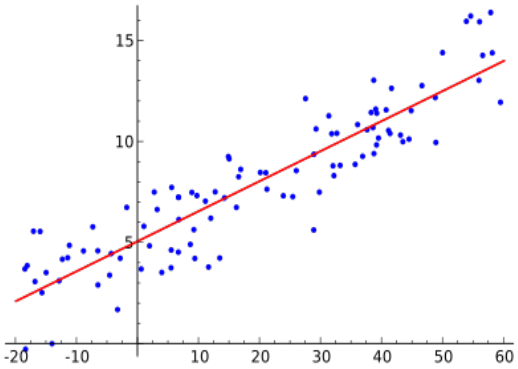

Figure 0.1: Linear Regression, source: Wikipedia/Google images

- Linear Regression is one of the most basic models which are used to predict future values of the dependent variable from the independent variables
- In our case the logerror which we had to predict was the dependent variable and the properties of each house given to us were the independent variables
- Using some of these independent variable and their corresponding log error, the linear regression model just plots a linear graph that best captures the data.
- It then calculates the mean error and r2 values to see how far each value is from the linear plot and then uses these error to more accurately predict the dependent variable(logerror)
- For linear regression I have used most of the properties given in the properties csv, just dropped a few that seemed unnecessary based on the correlation observed like propertyzoningdesc, propertycountylandusecode.
- Mean squared error : 0.03
- R2 score : 0.00

Random Forest Regression

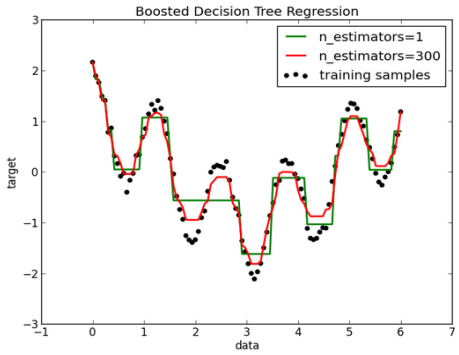

Figure 0.2: Random Forest Regression, source: Quora answers/Google images

- The essence of Random Forest Regression is decision trees. It captures the mean predictions of individual decision tress and thus doesn't let the decision trees be overfitting but infact gives a more precisive way to use a multitude of features and build decision tress on them
- It is specially useful when we are dealing with various missing values like in our dataset, this is one of the main reasons why I chose to use random forest regression
- With more trees in the forest random forest predictions are even better, so I have tried it with initially only 3 trees to then 50 trees and as expected 50 trees gives a lesser error
- I have used calculatedfinishedsquarefeet, fips, taxamount, structuretaxvaluedollarcnt as the independent variables to calculate the prediction. My idea of using these was the amount of correlation and also while trying to drop a few features I noticed these made a lot of difference to the final result
- Mean squared error: 0.00
- R2 value: 0.84

Analysis

- As observed from the above mean squared error values and r2 values, it shows us that random forest regression has definitely performed better then linear regression
- The mean squared error is lesser in Random Forest which means that its predicted is closer to the actual and the r2 values depend on the current set of features used for the prediction

Prediction Experience

- It was very interesting to see how certain important features made so much difference to the mean squared error and r2 value. So playing around trying to figure out the right set of values was interesting
- Another interesting thing was the closeness in kaggle scores, although so many of my prediction entries were very close to each other the ranks differed by a lot

Kaggle Evaluation

- My best score on Kaggle is 0.0648938 which was ranked 1997 at the time of submitting it and it then dynamically changed as per other submission, currently my best score ranks at 2018
- As mentioned above, the prediction scores were very close like 0.0648938, 0.0649105 etc