Assignment 2 Report

Rohan Vaish 111447435 (rvaish@cs.stonybrook.edu)

Multiple Linear Regression

How does Multiple Linear Regression work?

Multiple linear regression aims to map a linear relationship between two or more variables and a target variable by fitting a linear equation to the known dataset. Each value of an independent variable corresponds to some value of dependent variable.

If the relationship is as follows:

$Y = c_0 + c_1 X$ (where c_k is a constant coefficient) (1.1)

then X is the independent variable on which Y is dependent and this is known as simple linear regression. Y is known as the target or outcome variable and X is the predictor variable

If the relationship follows:

$Y=c_0 + c_1X_1 + c_2X_2 + c_3X_3 \ldots.+c_nX_n$ (where c_k is a constant coefficient) (1.2)

then X_k are the independent variables on which Y is dependent and this is known as multiple linear regression.

In our assignment, we had to predict the *logerror* for Zillow challenge which is defined as

logerror=log(Zestimate)−log(SalePrice) (1.3)

Using correlation coefficient matrix and basic understanding, I could identify 24 parameters of relevance to be used for multiple linear regression model-

$\mathbf{X_k}$=*['bathroomcnt','bedroomcnt','buildingqualitytypeid','calculatedfinishedsquarefeet','finishedsquarefeet12','finishedsquarefeet15','fips','garagecarcnt','garagetotalsqft','latitude','longitude','lotsizesquarefeet','regionidcity','regionidcounty','regionidneighborhood','regionidzip','roomcnt','unitcnt','yearbuilt','numberofstories','structuretaxvaluedollarcnt','taxvaluedollarcnt','landtaxvaluedollarcnt','taxamount']*

Using the ***"train_2016_v2.csv"*** and dividing this data into 2:1 ratio of training data: test data, the model was trained and tested. The predicted *Y_pred* was then compared to the actual *Y_test* and gave the following observations:

Multiple linear regression devises a linear line by calculating the least square distance and predicts the values by minimizing the squared sum of vertical displacements from each value to the line.

Evaluation of Multiple Linear Regression model

How well a model does can be evaluated in multiple ways. One of the ways is evaluating the ***Y_pred*** against ***Y_test*** with the help of ***mean squared error*** and ***variance score***. Given the list of 24 normalized parameters chosen for this model the following evaluations were obtained

Mean squared error ***LR***: 0.025694808 *Variance score* ***LR***: 0.003434294

Mean squared error measures the average of the squares of the errors and the lesser this value, the better the prediction model proves to be. It is always non-negative, and values closer to zero are better. In comparison to the other models used such as K Nearest Neighbours, Random Forest and AdaBoost, the Multiple linear regression model gave the best results on the train and test data. The following observation were made on the other models

Mean squared error ***KNN:*** 0.045762832

Variance score ***KNN***: -0.774898220

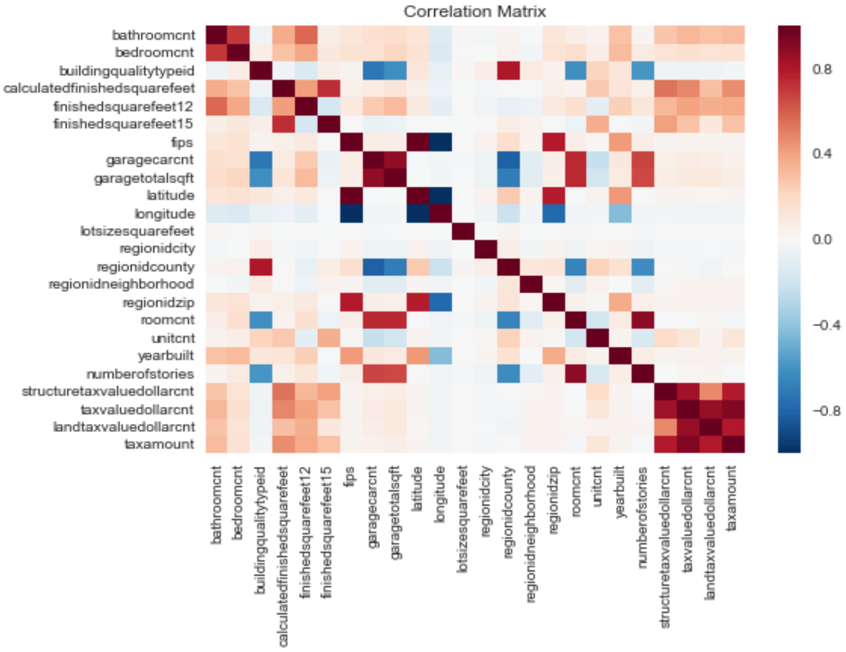
Mean squared error ***Random Forest***: 0.025703231

Variance score ***Random Forest***: 0.003107617

Mean squared error ***AdaBoost***: 0.048003035

Variance score ***AdaBoost***: -0.861783839

Using this data, the it was proven that multiple linear regression model was the best in predicting the ***logerror*** which in turn tells it was the best in predicting the price of properties for the Zillow challenge. The high correlation (visible from the heatmap) of the 24 parameters also helped these models in better prediction of the target variable.



Interesting Experiences and Surprises

- Personally, I got to learn a lot about handling such a large data of 2.7m rows
- Studying and cleaning of data was a big task while deciding which normalization to go for or how to handle missing data (by filling with 0 or mean)
- Given the superiority of AdaBoost over Linear regression in general, seeing a better performance from linear regression than KNN, AdaBoost, Random Forest was a big surprise
- While Multiple linear regression proved to be best with the training data, Random forest gave the best result in the Zillow challenge
- The scores were as follows:
    - Random Forest- 0.0649976
    - LR- 0.0652012
    - KNN- 0.1027224
    - AdaBoost- 0.1125407
- The final rank was ***2238***