**Springboard Capstone Project 1 – Predicting log-error of Zillow’s housing estimation**

**Inference statistics report**

The problem statement of this project is to predict the log error (y-value in the dataset) between the housing price estimation by Zillow and the actual sales price based on a list of features (x-value with features in the data set) given by Zillow. After data wrangling on null values and removing the features with 97% of null values, the following features are obtained represented by the correlation heat map. This correlation heat map gives us a high level of understanding on correlations between features, especially between features and log-error.



**Log-error (y-column in data set) analysis**

Basically, ***log-error is NOT correlated with all features in the data-set*** according to the above heat-map. The log-error is not to be predicted simply based on the correlations of data set features. Next we plot on the distribution of log-error.



Log-error distribution is nearly ***normally distributed with mean 0.011457 and standard deviation 0.161079***.

We further analyze log-errors by creating 10,000 bootstraps of log-error means and observe the distribution which also follows normal distribution with ***very small standard mean error for log-error: 1.7843 x 10-6*** and ***very small*** ***standard deviation of 10,000 bootstrap replicates of log-error mean: 0.000538912211383***.



**Feature (x-columns in data set) analysis**

By common sense, we can regard the following features are related each other:

|  |  |
| --- | --- |
| **Features** | **Correlation** |
| roomcnt (Number of rooms) vs.  numberofstories (Number of stories) | 0.6 to 0.8 |
| roomcnt vs.  fireplacecnt (Number of fire places) | 0.6 to 0.8 |
| garagecarcnt (Number of garages) vs.  garagetotalsqft (Total square feet of all garages) | Around 0.8 |
| taxvaluedollarcnt (Total tax assessed value of the property) vs. taxamount (Total property tax assessed) | Approaching 1.0 |
| bathroomcnt (Number of bathrooms) vs.  bedroomcnt (Number of bedrooms) | 0.6 to 0.8 |

Next we analyze correlations between features using statistical inference method as follows:

Assume that the observed correlation between two features may be just by chance. We set the null hypothesis H0: The true coefficient is equal to 0.

Permute one feature 10,000 times but leave another feature fixed. This simulates the hypothesis that they are totally independent of each other. For each permutation, compute the Pearson correlation coefficient and assess how many of the permutation replicates have a Pearson correlation coefficient greater than the observed one i.e. p-value of 0. Very low p-value means there is never a single replicate in the 10,000 permutations having a Pearson correlation greater than the observed one. Therefore, low p-value rejects H0 and conclude that the two feature are dependent of each other.

Next regression analysis of the two features is done.

**Analysis on number of garages vs. number of rooms**

Observed correlation on number of garages vs. number of rooms: 0.701320996434

p-value of the above null hypothesis: 0 (They are dependent of each other.)

Regression analysis

Range for slope of Number of cars in garage / Number of rooms with 95% confidence level: [ 0.22799573 0.23075117]

Range for intercept of Number of cars in garage / Number of rooms with 95% confidence level: [ 0.25675492 0.26640794]

