Springboard Data Science Career Track

Capstone Project 1: Machine Learning
Zillow Home Valuation Prediction

Introduction

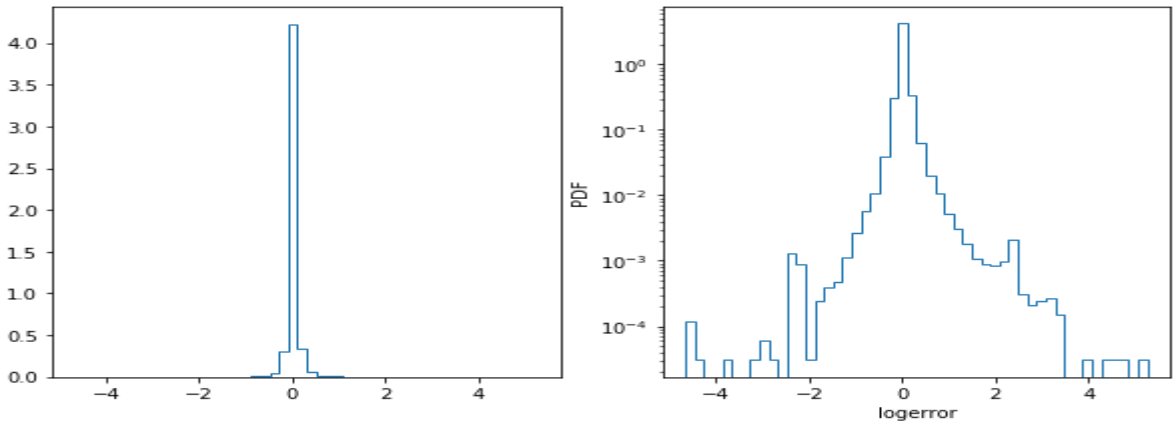
After cleaning and analysing the data we begin to build a prediction model. The target which we are trying to predict is 'logerror,' the log of the error from Zillow's own prediction algorithm. This is a continuous variable so we will need some kind of regression

First Models

We start with a basic linear regression to get some initial information on how to fit a more in depth model. These initial attempts can be found in the notebooks Linear_Regression.ipynb and Linear_Regression_pt2.ipynb.

Initial Difficulties

The data for 'logerror' is nearly entirely closer to 0, thus linear regressions will be very close to the line $y = 0$. However, this means that we will miss many of the points which are



further from 0. We can see the distribution of 'logerror' in normal scale to show how prevalent 0 is, and also in log scale to the structure of the data not near 0.

A fitted linear regression on this data gives an r2 score of only 0.004 and a residuals plot, which should be randomly distributed around 0 is displayed below.

We did not expect a linear regression to be a very good model, but this is much worse than expected.

Even some basic polynomial regressions do not do a much better job with similar r2 scores.

Reactions

The major reaction to this was to go back to data cleaning and rework the missing data using K-nearest-neighbours to guess at missing values. Additionally, gradient boosting was used to determine which features were the most useful

Final Models

Using what was learned in earlier models, I decided to separate ‘logerror’ into bins and use a model that begins as a classifier to get close to the ‘logerror’ value and then uses a ridge regression in each bin separately. This model is found in Model.ipynb.