Springboard Data Science Career Track

Capstone Project 1: Progress Report
Zillow Home Valuation Prediction

Introduction

Buying a house is a massive investment and carry with it a great deal of risk. Having accurate models which forecast home values based on empirical data can help both home buyers and financial institutions make sound decisions.

We will analyse a dataset of real estate transactions provided by Zillow and use it to build and train a model which predicts the error of Zillows own prediction algorithm.

Data wrangling and cleaning

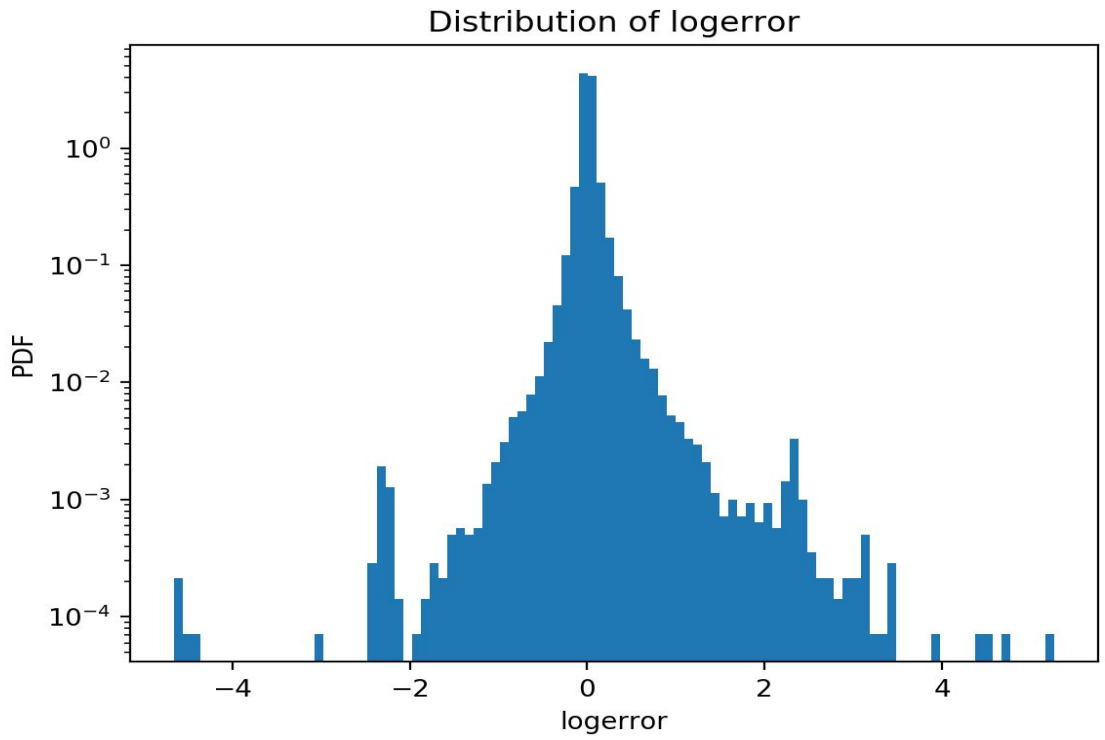
The dataset immediately provided two related problems. The first is that the dataset included entries which Zillow would use to evaluate the model. These values did not have a target or dependant variable associated with it and thus are useless to us. The other problem is that the dataset was quite large and unwieldy.

The solution to this was to use SQLalchemy to create a local database chunk by chunk and use SQL commands to filter and create a new csv with only those entries which have target values. The resulting csv was small enough to be easily handled by a personal computer.

The subsequent steps were fairly normal. Dropping entries and features that were mostly empty. Though some of these empty values were actually meant to represent 0 or FALSE values, so this was corrected.

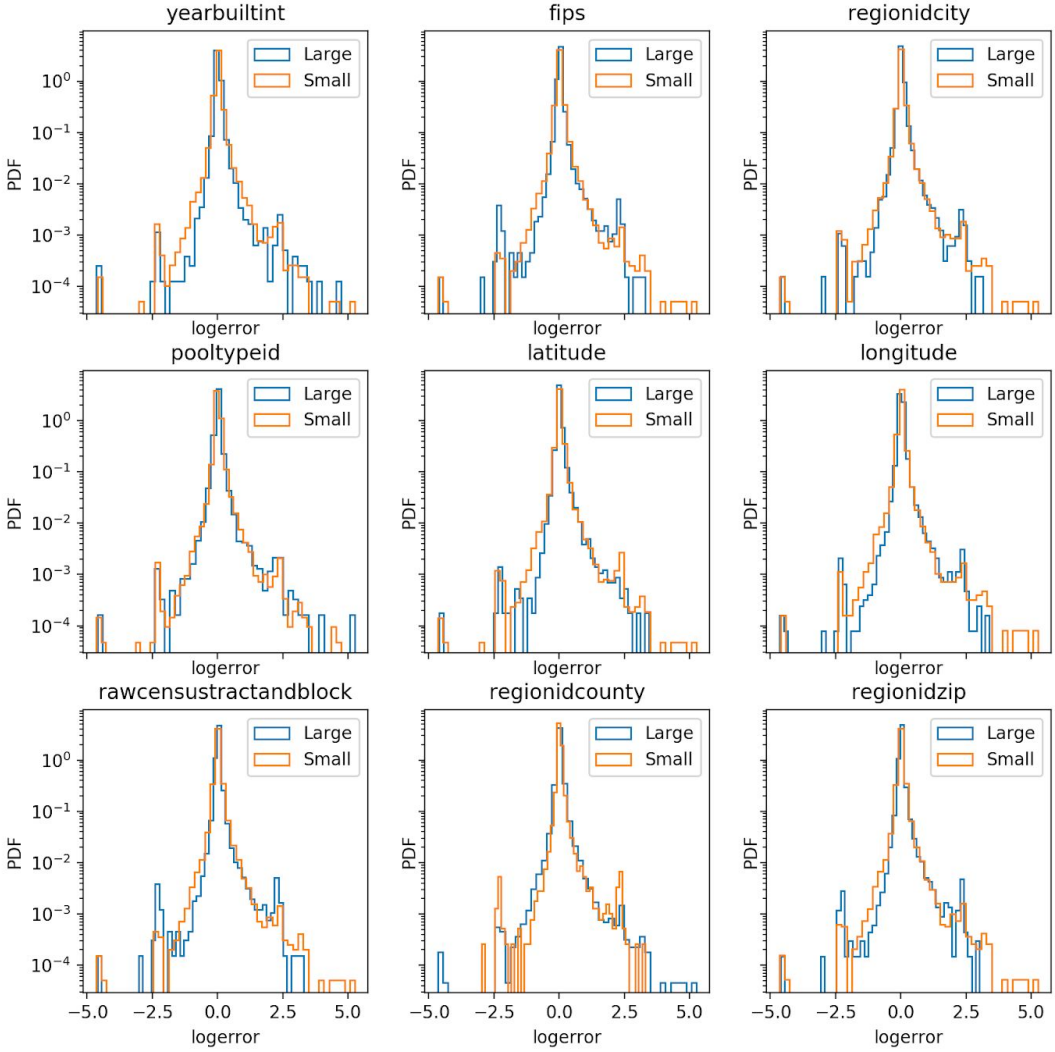
Exploratory Data Analysis

The next step was too look at different features and see if we could find any patterns. By looking at a histogram for the target variable ‘logerror’ we see that there are two clusters at around 2 and -2.



So we decided to look for features which may be have an outsized effect on the dataset causing these clusters. In doing so we identified 9 features which may have a different distribution for those values in the clusters. These were ‘yearbuiltin,’ ‘fips,’ ‘regionidcity,’ ‘pooltypeid,’ ‘latitude,’ ‘longitude,’ rawcensustractandblock,’ ‘regionidcounty’ and ‘regionidzip.’ We then looked for a cutoff point in each feature which would separate the values into a group that included the clusters and a group that did not.

We then performed a two sample bootstrapping test on each feature to verify that there was a difference in the distribution between the portion of the entries which included the clusters in the logerror and the others which did not.


yearbuiltint
fips
regionidcity
pooltypeid
latitude
longitude
rawcensustractandblock
regionidcounty
regionidzip
Large
Small
PDF
logerror