Capstone Project

Machine Learning Engineer Nanodegree

Wenting Rohwer
September 29th 2017

Project Overview

In this project, I will participate the Kaggle Zillow Home Value Prediction and will develop an algorithm that makes predictions about the future sale prices of homes.

Problem Statement

A home is often the largest and most expensive purchase a person makes in his or her lifetime. Ensuring homeowners have a trusted way to monitor this asset is incredibly important. The Zestimate was created to give consumers as much information as possible about homes and the housing market, marking the first time consumers had access to this type of home value information at no cost.

"Zestimates" are estimated home values based on 7.5 million statistical and machine learning models that analyze hundreds of data points on each property. And, by continually improving the median margin of error (from 14% at the onset to 5% today), Zillow has since become established as one of the largest, most trusted marketplaces for real estate information in the U.S. and a leading example of impactful machine learning.

Datasets and Inputs

Train/Test split

- I will be using a full list of real estate properties in three counties (Los Angeles, Orange and Ventura, California) data in 2016.
- The train data has all the transactions before October 15, 2016, plus some of the transactions after October 15, 2016.
- The test data in the public leaderboard has the rest of the transactions between October 15 and December 31, 2016.
- The rest of the test data, which is used for calculating the private leaderboard, is all the properties in October 15, 2017, to December 15, 2017. This period is called the "sales tracking period", during which we will not be taking any submissions.
- I will predict 6 time points for all properties: October 2016 (201610), November 2016 (201611), December 2016 (201612), October 2017 (201710), November 2017 (201711), and December 2017 (201712).
- Not all the properties are sold in each time period. If a property was not sold in a certain time period, that particular row will be ignored when calculating your score.

- If a property is sold multiple times within 31 days, we take the first reasonable value as the ground truth. By "reasonable", we mean if the data seems wrong, we will take the transaction that has a value that makes more sense.

File descriptions

- properties_2016.csv - all the properties with their home features for 2016. Note: Some 2017 new properties don't have any data yet except for their parcelid's. Those data points should be populated when properties_2017.csv is available.
- properties_2017.csv - all the properties with their home features for 2017 (will be available on 10/2/2017)
- train_2016.csv - the training set with transactions from 1/1/2016 to 12/31/2016
- train_2017.csv - the training set with transactions from 1/1/2017 to 9/15/2017 (will be available on 10/2/2017)
- sample_submission.csv - a sample submission file in the correct format

Solution Statement

I am planning on using different models to see which one gives the best predict. The models I will be using includes: XGBoost, LightGBM, and OLS and Keras NN

Benchmark Model

There is no Benchmark Model as the project is making prediction on how well the z-estimate predicts. The measurement is the difference between predicted log errors and the actual log errors.

Evaluation Metrics

Results are evaluated on Mean Absolute Error between the predicted log error and the actual log error. The log error is defined as

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

and it is recorded in the transactions training data. If a transaction didn't happen for a property during that period of time, that row is ignored and not counted in the calculation of MAE.