Zillow Prize: Zillow’s Home Value Prediction (Zestimate)

Deexith Mysore Nagaraj

INDEX

PROBLEM DESCRIPTION

Rules and Guidelines:

The Zillow Prize contest competition, sponsored by Zillow, Inc. ("Sponsor") is open to all individuals over the age of 18 at the time of entry. The competition will contain two rounds, one public and one private. Each round will have separate datasets, submission deadlines and instructions on how to participate. The instructions on how to participate in each round are listed below. Capitalized terms used but not defined herein have the meanings assigned to them in the Zillow Prize competition Official Rules.

OVERALL COMPETITION OVERVIEW:

First Round: May 24, 2017 – January 17, 2018 Pacific Time ("PT"). Total Prizes: $50,000

Second Round: February 1, 2018 – January 15, 2019 PT. Total Prizes: $1,150,000

Overview:

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

$$\text{Logerror} \quad = \quad \log(\text{Zestimate}) - \log(\text{SalePrice})$$

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

Submission File

For each property (unique parcelid), you must predict a log error for each time point. You should be predicting 6 timepoints: October 2016 (201610), November 2016 (201611), December 2016 (201612), October 2017 (201710), November 2017 (201711), and December 2017 (201712). The file should contain a header and have the following format:

```
ParcelId,201610,201611,201612,201710,201711,201712

10754147,0.1234,1.2234, -1.3012,1.4012,0.8642-3.1412

10759547,0,0,0,0,0,0, etc.
```

Note that the actual log errors are accurate the 4th decimal places, so you can adjust your decimal formats to limit the size of your submission file.

DATA 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 parcel id'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 (released 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 (released on 10/2/2017)

sample_submission.csv - a sample submission file in the correct format

Approach:

- Data is processed by removing the unwanted features or the predictors.
- Sklearn library has random-forest-regressor and that has been used here to build the model.
- After playing with the hyperparameters, the model was built set the produce the best result.
- Test data is the available submission sample file provided.
- The built model is then used to predict the value and log error is calculated.

User Manual:

Requirements:

- Python3.6
- Sklearn, pandas, numpy
- Jupyter Notebook

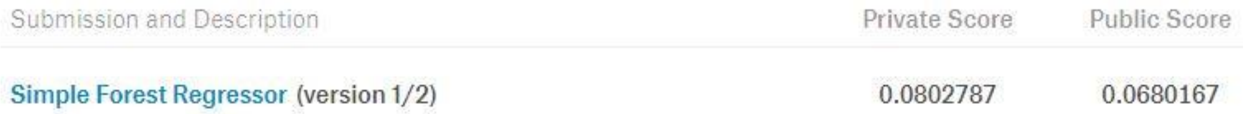
Run Procedure:

- Open the file in Jupyter notebook and install necessary libraries.
- Once all the dependencies are cleared, select Kernel and Run All.

- The predicted value is written on to the csv file “PredictedValue”

Result:

Below are the results observed.



Submission and Description	Private Score	Public Score
Simple Forest Regressor (version 1/2)	0.0802787	0.0680167

Screen Shot of the score