Project-5

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

CS514 Applied Artificial Intelligence

Problem Statement

Zillow's Zestimate home valuation has shaken up the U.S. real estate industry since first released 11 years ago. 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.

In this competition, participants will develop an algorithm that makes predictions about the future sale prices of homes. The contest is structured into two rounds, the qualifying round which opens May 24, 2017 and the private round for the 100 top qualifying teams that opens on Feb 1st, 2018. In the qualifying round, you'll be building a model to improve the Zestimate residual error. In the final round, you'll build a home valuation algorithm from the ground up, using external data sources to help engineer new features that give your model an edge over the competition.

Data Description

- You are provided with 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.
- You are asked to 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).

File Description

- properties_2017.csv - all the properties with their home features for 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
- sample_submission.csv - a sample submission file in the correct format

Methodology Used

- The data is imported and preprocessed by filling in all the missing values, remove outliers converting the datatype of the columns to required format for training purpose.
- Concatenated the two training datasets, and merged the training and properties datasets.
- The combined data is split into training and test data. Features that are not important are dropped.
- Gradient Boost Regressor is used for training the model and mean absolute error is predicted for the test data.
- Tried using GridSearchCV to get the important features and standardizing the training data.

Steps used to run the program

- Install Python 3.x
- Download all the necessary packages
- Add “Zillow.py”, training, properties and sample submission files in the same directory.
- Execute “Zillow.py”. It will produce “submission.csv.”
- Submit the prediction file to Kaggle for evaluation and check the score.

Result

Private score : 0.0755891
Public score: 0.0646444