PROJECT 5

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

Megamind | 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.

"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.

Zillow Prize, a competition with a one million dollar grand prize, is challenging the data science community to help push the accuracy of the Zestimate even further. Winning algorithms stand to impact the home values of 110M homes across the U.S.

Data Description :

In this competition, Zillow is asking you to predict the log-error between their Zestimate and the actual sale price, given all the features of a home. The log error is defined as:

Logerror = log (Zestimate) - log (Saleprice)

and it is recorded in the transactions file train.csv. In this competition, you are going to predict the logerror for the months in Fall 2017. Since all the real estate transactions in the U.S. are publicly available, we will close the competition (no longer accepting submissions) before the evaluation period begins.

File Description :

- 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 (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

Data Fields :

- id - an anonymous id unique to a property
- feat_1, feat_2, ..., feat_n - the various features of a properties
- target - the log error of a properties

Methodologies Used :

- The data is imported and preprocessed by dropping all columns that have null i.e. rows and labels before training.
- Then, GridSearch CV was used to tune the hyper parameters using the validation data set.
- Finally, those tuned hyper parameters were used for the parameter list of Light GBM (Gradient Boosting Model) method is used for training and testing the data.
- Also tried ensemble of classifiers, but score was not good enough.

STEPS TO RUN THE PROGRAM:

Pre-Requisites:

1. Python 3.x - Anaconda (download here)
2. Sklearn, Pandas, LightGBM and other operator modules.
3. Note: To install Light GBM in Anaconda, run “conda install -c conda-forge lightgbm ”.

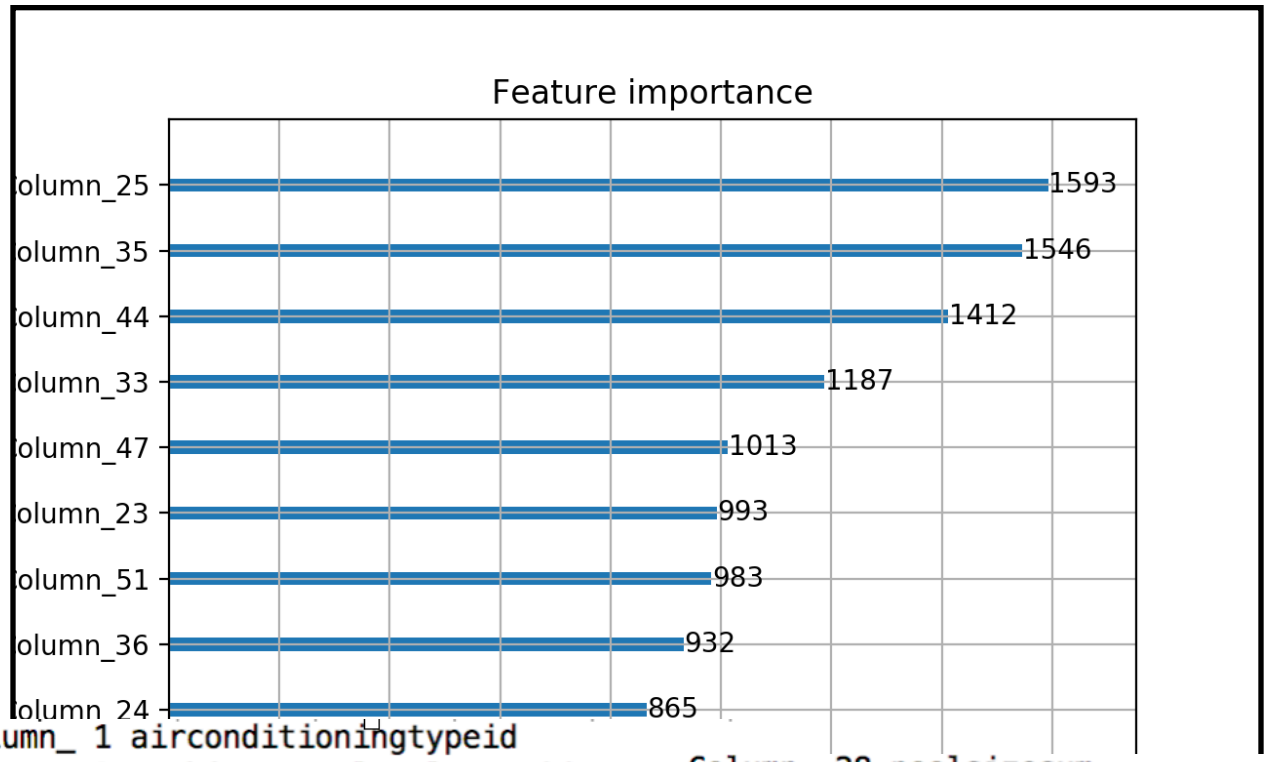
Instructions:

1. Add Megamind.py, sample_submission.csv, train_2016_v2.csv and properties_2016.csv in the same folder “input”.
2. Navigate to Megamind/input/ and execute Megamind.py.
3. It will produce a Results.csv file in the same folder “input".
4. The prediction file can be submitted to Kaggle for evaluation and check the score.

Results:

The best public score my code got was 0.0650130 and private was 0.0758792. The best score overall public score was 0.0631885 and private was 0.0740861.

Graphs for VISUALISATION: TOP FEATURES



Column_ 1 airconditioningtypeid
Column_ 2 architecturalstyletypeid
Column_ 3 basementsqft
Column_ 4 bathroomcnt
Column_ 5 bedroomcnt
Column_ 6 buildingclasstypeid
Column_ 7 buildingqualitytypeid
Column_ 8 calculatedbathnbr
Column_ 9 decktypeid
Column_ 10 finishedfloor1squarefeet
Column_ 11 calculatedfinishedsquarefeet
Column_ 12 finishedsquarefeet12
Column_ 13 finishedsquarefeet13
Column_ 14 finishedsquarefeet15
Column_ 15 finishedsquarefeet50
Column_ 16 finishedsquarefeet6
Column_ 17 fips
Column_ 18 fireplacecnt
Column_ 19 fullbathcnt
Column_ 20 garagecarcnt
Column_ 21 garagetotalsqft
Column_ 22 hashottuborspa
Column_ 23 heatingorsystemtypeid
Column_ 24 latitude
Column_ 25 longitude
Column_ 26 lotsizesquarefeet
Column_ 27 poolcnt
Column_ 28 poolsizesum
Column_ 29 pooltypeid10
Column_ 30 pooltypeid2
Column_ 31 pooltypeid7
Column_ 32 propertylandusetypeid
Column_ 33 rawcensustractandblock
Column_ 34 regionidcity
Column_ 35 regionidcounty
Column_ 36 regionidneighborhood
Column_ 37 regionidzip
Column_ 38 roomcnt
Column_ 39 storytypeid
Column_ 40 threequarterbathnbr
Column_ 41 typeconstructiontypeid
Column_ 42 unitcnt
Column_ 43 yardbuildingsqft17
Column_ 44 yardbuildingsqft26
Column_ 45 yearbuilt
Column_ 46 numberofstories
Column_ 47 fireplaceflag
Column_ 48 structuretaxvaluedollarcnt
Column_ 49 taxvaluedollarcnt
Column_ 50 assessmentyear
Column_ 51 landtaxvaluedollarcnt
Column_ 52 taxamount
Column_ 53 taxdelinquencyflag

THANK YOU