Applied Artificial Intelligence Project -5

Nickname: ZestBlister

Domain: Zillow's Zestimate home valuation

Background: 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.

Steps to run the code: (Assuming you are a user on Kaggle)

- Go to https://www.kaggle.com/c/zillow-prize-1/kernels
- Add a New Kernel.
- Copy the code from the attached script and run it (or upload to run).
- You can choose the models you want to see the results(also in the table below) for by un-commenting the relevant code. As default the script runs LightGBM which gave me the best results.
- After execution, go back and in the Output tab select Submit to Competition to see the scoring results.

Model/Techniques:	Test Scoring as per Kaggle evaluation
RandomForestRegressor (with CrossValidation)	0.0775136
LightGBM (Trained with a validation set)	**0.0647271**
ExtraTreesRegressor (with Cross Validation)	0.0822319
MLPRegressor (with Cross Validation)	0.0652381
BayesianRidge (with Cross Validation)	0.0650652
LinearRegression (with Cross Validation)	0.0651190

Conclusion: Basic machine learning regression techniques performed quite well and with better feature engineering the results can be improved. Light Gradient Boosting (LightGBM) generally is preferred for large data sets since it is efficient (in both memory usage and time) and gives better accuracy. Its results could also be improved by better parameter tuning and feature engineering. Also using just the 2016 dataset rather than the 2017 or a merged dataset gave better results for all the techniques used and hence only 2016 dataset was used to train the models.