Model prediction on house value

The goal of the project is to build a model to improve the Zestimate log-error between Zestimate price and actual sale price. Increasing the model performance of predicting the market values of the properties will benefit homeowners, investors, and realtors. In this project, we will investigate the datasets to value the importance of contributing factors for prediction, and then will apply different machine learning methods and evaluate the model which provides the best capacity.

1. Data
   Zillow Prize: Zillow’s Home Value Prediction (Zestimate) is Open source data from Kaggle https://www.kaggle.com/c/zillow-prize-1/overview). The overview of datasets can be found on the website. To be brief, it provided two sets of training sets and properties datasets from 2016, and 2017 separately. And a test (submission) file format is also provided.
   The goal of this project is to predict log error in 6 time points. The feature description is provided as a dictionary.

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

2. Explory data analysis:
   a. There are many missing features that have a relatively high percentage. In the Catboost model, we exclude 98% of the missing percentage from the features.
   b. 15 features were excluded initially. And the features that only have one unique value were excluded.

3. Model: Catboost
   a. Data preprocessing:
   This algorithm is fast, scalable, high performance Gradient Boosting on Decision Tree. It can be used for ranking,

classification, regression, and other machine learning tasks. This project used regression to solve the problem.

Here, different datasets were merged in the year, such as properties and train data. Later, train data includes the whole datasets from 2016 and 2017.

b. Transcationdate is informative and was transformed as datetime features: including year, quarter, month and day compared to 2016.
c. Based on missing value and unique value screening, we kept 42 features for modelling. Then, the categorical variables were defined from the whole training features.
d. Imputation: instead of traditional imputation, we fillna with ‘None’ string, and then all categorical variables have to be converted to categorical or string type (essential!!)
e. After defining the training and test data, the model was instantiated and run with experienced parameters.
f. Submission is the output of the Y_predction at different time points at each ParcelIID.
g. Score in Zillow is described as Zillow MAE. The public leaderboard score is 0.06318 and our prediction shown is the score is as small as 0.054393.

4. Discussion

This project uses Catboost as it is an innovative algorithm for processing categorical features. No need to preprocess features on your own. The implementation of ordered boosting is an alternative to the classic boosting algorithm. Future work should consider hyperparameter tuning with Grid and Random Search.