Capstone Project Proposal - Machine Learning Engineer Nanodegree

Lei Pan
November 11th, 2017

Zillow’s Home Value Prediction - “Zestimate”

Domain background

- Zillow created “Zestimate” which gives customers a lot of information about homes and housing markets at no cost by using publicly available data.
- 7.5 million statistical and machine learning models that analyze hundreds of data points on each property are used by Zillow to create and improve “Zestimate”. They improved median margin of error from 14% to 5%. Zillow announced a Kaggle competition to improve the accuracy of “Zestimate” even further.
- Zillow competition has two rounds. The first round is to build a model to predict Zillow residual error. The final round is to build a home evaluation algorithm from ground up using all external data. My project will focus on the first round of the competition. The goal of capstone project is to build a model to improve Zillow residual error.
- This is a very typical supervised machine learning problem, because supervised learning algorithms learns and analyzes labeled training data and then generates function to predict output. Zillow gave the datasets of log error between Zestimate price and actual price for both 2016 and 2017 which are labeled data as well as Zillow asked for a prediction for log error. Similar machine learning tasks are weather apps predict the temperature for a given time and spamming emails prediction based on prior spamming information.

Problem Statement

- A machine learning computer program is said to learn from experience E with respect to some class of task T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. In this capstone project:
    - P is Mean absolute error of predicted log error and actual log error.
    - T is The log error prediction task.
    - E is The process of the algorithm examining a large amount of historical data of log error.

Datasets and Inputs

- Training sets
    - Train_2016.csv. It contains transactions from 1/1/2016 to 12/31/2016
- Testing sets
    - Train_2017.csv. It contains transactions from 1/1/2017 to 9/15/2017
- Other data sets
    - properties_2016.csv - properties for the home features for 2016.
    - properties_2017.csv - properties for the home features for 2017.
- You can find data here: https://www.kaggle.com/c/zillow-prize-1/data
- Analysis of the training dataset:
    - Shape of Train_2016.csv is (90811,3)
    - Shape of properties_2016.csv is (2985217,58)
    - Since we have 90811 rows in train dataset but 2985217 rows in properties dataset, we need to merge two files to do our data analysis.

Solution statement

- Since the goal is to predict the log error between Zestimate and actual price and we have all the training and testing dataset for it, this is a very clear supervised learning problem for me. Based on the research I've done regarding the supervised learning algorithms, lightGBM and XGBoosting are clear winners for this problem statement in terms of accuracy and performance. First, I will clean up dataset including dealing with null values and missing values as well as converting non-numerical data to numerical data and remove outliers. Then I will do feature selections to find most important features. Then I will build lightGBM and XGBoosting models and tune parameters to get the predicted results and mean absolute errors. Finally, I will ensemble two models together to get the best results.

Benchmark model

- During my research, I found one very good baseline lightGBM model which has a very good model with a very less mean absolute eror - 0.06487. I can tune and refine my first lightGBM model based on this model. This is the model I refer to: https://www.kaggle.com/guolinke/simple-lightgbm-starter-lb-0-06487

Evaluation metrics

- Models are evaluated on Mean Absolute Error between the predicted log error and the actual log error. The log error is logerror=log(Zestimate)−log(SalePrice)
- If I can build a model with a MAE that's less than 0.06487(base model), then the model passes evaluation.

- The reason I chose this evaluation metrics is that 1) this is provided by Zillow to evaluate actual competition. 2)Since every Zillow competition participant uses and publishes this evaluation metrics, I can compare my MAE with all the MAEs for Zillow competition participants' models to get a sense how well I am doing with all the professionals around the world.

Outline of the project design

1. I will check missing values, null values, and empty values for all the features.
2. Feature selection and feature engineering will be done based on data visualization and correlation analysis I will perform on the dataset.
3. I will try out two algorithms 1). LightGBM model. 2). XGBoosting model.
4. For LightGBM Model, I will use the latest Model from lightGBM.library.https://lightgbm.readthedocs.io
5. For XGBoosting model, I will use the latest XGBoosting library. https://xgboost.readthedocs.io/en/latest/
6. To boot performance, I will cross validate and use hyperparameter optimization.
7. Both models will be tested.
8. I will combine those two models together for the best results.