Capstone Project Proposal - Machine Learning Engineer Nanodegree

Lei Pan

October 25th, 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.
- First round of the competition is about using their existing data to push accuracy of "Zestimate" even further. Second round of the competition is that competitors can bring external data sets to improve the model. The scope of my capstone project is to finish the first round of this competition.

Problem Statement

- The goal of the project is to predict the log-error between the predicted prices from the model and actual sales price. Mean Absolute Error between the predicted log error and the actual log is used to evaluated the model I am going to develop. It is defined as follow:
  logerror=log(Zestimate)−log(SalePrice)logerror=log(Zestimate)−log(SalePrice)
  https://www.kaggle.com/c/zillow-prize-1#evaluation

Datasets and Inputs

- train_2016.csv - training set. It has transactions from 1/1/2016 to 12/31/2016
- train_2017.csv - training set. It has transactions from 1/1/2017 to 9/15/2017
- properties_2016.csv - properties for the home features for 2016.
- properties_2017.csv - properties for the home features for 2017.
- sample_submission.csv - a sample submission file.
- https://www.kaggle.com/c/zillow-prize-1/data

Solution statement

- Since the goal is to predict the log-error between the predicted price and real price and we have all the training and testing dataset for it, this is a very clear supervised learning problem for me. Among all the supervised learning algorithms that I learned through nano degree course, Gradient Boosting model would be a good fit for this problem; because I tested and compared it with other algorithms on multiple supervised learning projects and it gave me the best result. In addition to Gradient Boosting model, another model seems very interesting to me is XGBoosting. This is a great chance to try this model out. In summary, I am going to use both Gradient Boosting model and XGBoosting model to solve the problem. After comparing the results from those two models, I will pick up the best model from this two models.

Benchmark model

- Since Zillow provides their residual errors as well as the their property data and between their estimate and the actual sale prices, I will use the existing Zestimate model as benchmark model.

Evaluation metrics

- The log-error between estimation price and the actual sale price will used to evaluate the model I am going to develop.
  Formula for the residual error:
  logerror=log(Zestimate)−log(SalePrice)logerror=log(Zestimate)−log(SalePrice)
- I will train the model on data from 2016 and test the model on data from 2017 using the formula above.

Outline of the project design

- First, I will check the data to see if we have missing values and decide how to handle the data.
- Then, I will do feature selection and feature engineering based on data visualization and correlation analysis I will perform on the dataset.
- I will try out two algorithms 1. Gradient Boosting Model. 2. XGBoosting model.
- For Gradient Boosting Model, I will use Gradient Boosting Model from python sklearn library.
- For XGBoosting model, I will use the latest XGBoosting library. https://xgboost.readthedocs.io/en/latest/
- To boot performance, I will cross validate and use hyperparameter optimization.
- In the end, I will test both models.
- I will pick up the winner for the best results.