Machine Learning Engineer Nanodegree

Capstone Proposal

Jinchuan Wei
Nov 18, 2018

Proposal

Domain Background

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. A method to estimate home price can give consumers as much information as possible about homes and the housing market. Technology driven real estate companies such as Zillow and Redfin provide solutions to estimate individual home values based on millions of home information nationwide.

Problem Statement

A machine learning model can be created to estimate individual home prices based on some specific home metrics such as floor plan, number of rooms, tax delinquency years, etc. The performance of created machine learning model can be evaluated on mean absolute error between the predicted log error and the actual log error. The log error is defined as

$$\text{logerror}=\log(\text{EstimatedPrice})-\log(\text{SalePrice})$$

Datasets and Inputs

Datasets from Zillow's Home Value Prediction Contest can be used as inputs for model training and testing in this project. The url link to the data set is https://www.kaggle.com/c/zillow-prize-1/data. A full list of real estate properties in three counties (Los Angeles, Orange and Ventura, California) data in 2016 is provided. The train data has all the transactions before October 15, 2016, plus some of the transactions after October 15, 2016. In the project transactions and real estate properties data are split into train and test dataset. Train dataset are used to train machine learning model, and test dataset to test on the performance of built machine learning model.

Solution Statement

Regression supervised machine models can be created to predict home values. A couple of ready to use machine learning models include ensemble methods, support vector machines, naïve bayes and decision trees. logerror mentioned in problem statement section can be used to tune model performance. Deep learning models such as convolutional neural networks can also be built to predict home values.

Benchmark Model

Zillow has a proprietary individual home value prediction solution Zestimate. The Zestimate's accuracy (https://www.zillow.com/zestimate/#acc) depends on location and availability of data in an area. Zillow's accuracy has a median error rate of 4.3%. This means half of the home values in the area are closer than the error percentage. For example, in Los Angeles-Long Beach-Anaheim, CA, Zestimate values for half of the homes are within 3.6% of the selling price, and half are off by more than 3.6%. The models to be built in this project are measured against the performance of Zestimate.

Evaluation Metrics

Since all transactions data are presented, model estimated home values are measured against real transaction price. Median error rate is calculated on the built model and measured against benchmark model.

Project Design

Data cleaning and transformation: the raw data provided by Zillow competition is not perfect. Before building model on the training data, data cleaning should be done first, i.e. remove outliers, feature scaling, detect null values, etc. The house properties contain some properties that may have little impact on the home values. Principal component analysis can be done on the raw data to reduce dimensionality and select most important features.

Multiple models can be built and tuned to predict properties values. Simple unsupervised learning models such as ensemble methods, support vector machines, naïve bayes, decision trees, random forest regression will be used. More advanced techniques include deep neural networks.

Models' performance will be tested against benchmark model, and the best model is selected to be the final solution.