Udacity Machine Learning Engineer Nanodegree Capstone Proposal

Kaggle-Zillow Prize: Zillow's Home Value Prediction (Zestimate)

Allan Yong, 27/06/2017

Project Overview

This project is based on the Kaggle competition described at: Zillow's Home Value Prediction and seeks to improve the accuracy of house valuations. The median margin of error has improved over the years from 14% to 5%. However, reducing the error further has a meaningful impact on potential homeowners as a home is usually their largest purchase. Zillow has kindly agreed to provide access to a treasure trove of home valuation data in the hopes of obtaining a better valuation models for homes.

Kaggle's competition is appealing as a project as it allows the data scientist to explore various machine learning techniques without focusing on data collection.

Problem Statement

The inputs would be in the form of structured data where 2985217 data points and 59 features such as number of bedrooms will fed into a regressor-type machine learning algorithm to generate predictions of the difference between actual house price and Zillow's prediction.

Datasets and Inputs

Zillow has provided full list of real estate properties in three counties (Los Angeles, Orange and Ventura, California) data in 2016. There will be **2985217** data points with **58** features that will be used to predict the target variable (difference between actual house price and Zillow's prediction). Some of these features will be categorical such as airconditioningtypeid and numerical such as calculatedfinishedsquarefeet.

Solution Statement

Various machine learning techniques will be applied to the dataset to predict the mean absolute error (MAE) between the predicted log error and the actual log error of the house price. The simplest technique to approach this problem would be to apply multiple regression. This method seeks to establish a relationship between the 2 or more independent variables (house size, distance from center of the city) and a single dependent variable (mean absolute error of the log of house prices).

A collection of decision tree models in Xgboost or Lightgbm will use the features of the training data to make models to fit the training data. These models will then be saved and used to make a prediction on the test sets. The models are saved in practice to avoid having to re-run the preliminary models again when included in an ensemble of models.

Later on a few of these models may be stacked together to combine the predictions other regressors to obtain a better prediction of the error. Stacking multiple models is expected to improve the score over the individual models if the models are sufficiently uncorrelated.

Evaluation Metrics

The project success will be evaluated on the score improvement over the benchmark model, as given in the competition and the mean absolute error between the predicted log error and the actual log error of the house price. A lower error would be preferred.

The MAE will provide a simple measure of the prediction error that disregards the sign of the error and doesn't over-emphasize outliers.

The prediction time as well as training time will be recorded compared to estimate/quantify the computational workload required in a production environment. These times will be used with the final scores to determine viability of the model

Benchmark

The benchmark model for this project has been supplied by Kaggle. However, the mean MAE could serve as a simple benchmark. Another benchmark could be a simple linear regression based on the actual data and the preprocessed data. The final model with be compared these simple models in order to judge its performance and quantify its improvements over the most simplistic model.

Project Design

1. **Data Exploration**

Simple statistics of the dataset will be calculated to obtain a rough idea of the mean, spread, completeness and quality of the data. Missing values and outliers will be dealt with accordingly. Data points may either be removed or scaled.

2. **Exploratory Visualization**

Multiple matplotlib, ggplot, seaborn plots will be generated explore the data to identify important features, gain some intuition on the problem and perhaps discard features which provide no information or which have insufficient data points.

3. **Data Preprocessing and feature engineering**

- Categorical data is transformed to numerical via one-hot encoding.
- Data is scaled to 0-1 with the Minmax transformation, or demeaned and scaled with its standard deviation (StandardScaler).
- Redundant features may be removed based on the least significant features from Xgboost and Lightgbm or the most correlated features.
- New features may be created from clusters or PCA.

4. **Model building and Selection**

- Several standard regressors such as the linear regression, ridge regression, RandomForestRegression, ExtraTreesRegression, NearestNeighborsRegression, GradientBoostedRegression such as Xgboost and Lightgbm will be applied to the problem.
- The hyper-parameters of the model will be tuned via Gridsearch to obtain the best prediction model.
- An ensemble of a few promising models will be generated to improve their performance.