Udacity Machine Learning Engineer Nanodegree
Capstone Proposal

Amrish Purohit June 25, 2018

Domain Background

Estimating price of a house is quite challenging and important process as housing prices are difficult to predict, and can be influenced by a very large quantity of factors. Everyone want better home in lease amount. This is lifetime decision and people spends hundreds of thousands of dollars to buy a house (at least in Los Angeles and major cities of California and US). To make such important decision, buyer want to make sure that fair price is placed on the property, particularly price of the house is not inflated.

Now days there are many real estate companies provide data and own define algorithm to determine best price of the house. One of such company is Zillow. Zillow has millions of data on homes across United States. Zillow has machine learning algorithm called "Zestimate". The Zestimate home value is Zillow's estimated market value for an individual home and is calculated for about 100 million homes nationwide. The Zestimate is calculated from public and user-submitted data, considering special features, location, and market conditions. More information about Zestimate can be obtained from here Zestimate.

Problem Statement

Based on data set provided on Kaggle, https://www.kaggle.com/c/zillow-prize-1, goal of the project is to predict price of new property going to sold in Los Angeles, Orange and Ventura county of California. Here price of home is evaluated based on features of home, it is a regression problem.

Datasets and Inputs

For this project, I will use data set provided by Zillow on Kaggle competition, https://www.kaggle.com/c/zillow-prize-1. The train data has all the transactions before October 15, 2016, plus some of the transactions after October 15, 2016 and test data in the public leaderboard has the rest of the transactions between October 15 and December 31, 2016. About 58 features provided in properties file like size, neighborhood, tax and location.

The data set will be used to train model to estimate price of home and prediction will be compared with Kaggle board.

Solution Statement

This project is classic example of regression problem. Initially I was planning to use deep learning technique to implement the algorithm as I am very impressed with deep learning during course of the program. As feature set is relative small in context of deep learning and comment I got from my connect program teacher, It may tend to over fitting on training set.

I will use tree based regression algorithms to implement the solution as they are easy to work, relatively fast trainers and robust to outliers and random noise present in data value. I will use decision tree regressor and variant or it. As decision tree is tend to overfit I will test training set with random forest and other Gradient boosting regressor like XGBoost and Light GBM

Benchmark Model

For the benchmarking, I will use simple linear regression model. I will take score of regression model as a base score and will implement and compare score of tree base regression models.

Evaluation Metrics

The evaluation metric that will be used in this project include Mean Absolute Error(MAE). MAE is defined as

$Logerror$ = log($Zestimate$) − log($SalePrice$).

it is recorded in the transactions training data. If a transaction didn't happen for a property during that period, that row is ignored and not counted in the calculation of MAE.

Project Design

1. Setup environment with python3 and other libraries like Numpy, Pandas, Matplotlib and Seaborn.
2. Exploratory Analysis of Data (EAD): Write up some sample code to read csv files provided in Kaggle data set. Plot data in histogram to understand distribution of the data. This will help to understand the feature and will provide a direction to move forward at feature selection.
3. Finding correlation between features: with the help of heatmaps and correlation matrix, I will study available feature sets to find correlation between the features. Understand correlation between single or multiple features to log error. This will help me create features out of features if required.
4. Data preprocessing and cleaning: Handle the missing values from observation and either remove the observation of provide mean values. Identify the outliers and remove them from the observation before fitting it in model.
5. Split data: Split cleaned and preprocessed data in train and test set.
6. Develop Initial models: I implement initial model of Random Forest Regressor and other Gradient Boosting model like XGBoost and light GBM.
7. Tuning Hyperparameter: Parameter technique that I learnt in course like Grid search, will be used to tune the parameters of the model with training data.
8. Model Evaluation and further tuning: Once done with parameter tuning, I will evaluate the model with test data. Based on the result/outcome of the test data, One of the above mentioned step will be repeated to improve the score of the model.

References:

1. https://www.zillow.com/zestimate
2. https://www.kaggle.com/c/zillow-prize-1
3. https://en.wikipedia.org/wiki/Gradient_boosting
4. https://en.wikipedia.org/wiki/Random_forest
5. https://en.wikipedia.org/wiki/Xgboost