SACRAMENTO STATE
COLLEGE OF ENGINEERING & COMPUTER SCIENCE

CSC 215-01 Artificial Intelligence (Fall 2018)

Final Project Proposal :
Home Truths : A housing price prediction model using advance regression techniques

PROJECT TYPE
A
(Combination of Option 2 and 3)

By
Akshada Rasam
Sac State ID - 219218000

Tushar Gandhi
Sac State ID – 219387208

Due, Tuesday, Oct 30, 2018, 11:59 p.m.

Project Description/Abstract:

Usually, House price index represents the summarized price changes of residential housing. While for a single family house price prediction, it needs more accurate method based on location, house type, size, build year, local amenities, and some other factors which could affect house demand and supply. As the famous Boston housing dataset is very limited, in this project, we plan on using our model on 3 different datasets namely The Boston Housing dataset, secondly , the Ames Housing dataset and finally the Singapore Housing Prices dataset. We would then apply a few. Of the tradition traditional regression models like Linear Regression, Logistic Regression, Polynomial Regression, Stepwise Regression, Ridge Regression, Lasso Regression etc. We would also implement the hybrid Lasso and Gradient boosting regression model to predict individual house price with reference to a paper "A Hybrid Regression Technique for House Prices Prediction". We will also perform a comparative study on feature selection using PCA, Univariate ANOVA f-test and apply boosting algorithms like LightGBM, XGBoost and Catboost and evaluate the performance using RMSE score.

Motivation:

In the real estate industry, when a house is to be sold or bought, the corporate world treats houses as a marketing product, tagging each house with its most beneficial features. Amongst the hundreds of features that each of these agents have, a layman would be familiar with only a few of them. Some of the obvious features of a house would be the total area of the house, number of bedrooms, number of bathrooms, locality or neighborhood, proximity to shopping, schools, hospitals etc. However, even if one house is exactly similar to other in all the obvious parameters yet there is a big difference in their selling price. Thus, it is very important to know the various features that impact the selling price. Also, The selling price of a house is important to evaluate the economical importance of a location.

Background:

Predicting housing prices has been a very popular problem when it comes to understanding the basic concepts of machine learning. There have been many papers that are published in this literature regarding houses of different cities and states around the world. We are referring to a hybrid regression technique that has been studied in the following paper :
Refer : https://ieeexplore-ieee-org.proxy.lib.csus.edu/stamp/stamp.jsp?tp=&arnumber=8289904

Data sets

1. *The Ames Housing dataset*
   https://www.kaggle.com/c/house-prices-advanced-regression-techniques
   An Ongoing competition on Kaggle.com. The Ames Housing dataset was compiled by Dean De Cock for use in data science education. It contains 2930 observations on 80 predictor variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous). These were used in the valuation of homes in Ames Iowa during the period 2006 until 2010.

2. *Boston housing dataset.*
   http://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_boston.html

3. *Singapore Housing Prices*
   https://www.kaggle.com/c/singapore-housing-prices
   There are 2 different datasets in this package, namely 'hdb' and 'private'. 'hdb' represents public housing in Singapore while 'private' represents private housing in Singapore. The dataset is considerably large, given that hdb contains 763766x15 and private contains approximately 461kx21 amount of data.

System/Algorithmic design:

The house price prediction is a Regression problem. Thus we are planning to train the model on a few traditional regression models like :

1. Linear Regression
2. Logistic Regression
3. Polynomial Regression
4. Stepwise Regression
5. Ridge Regression
6. Lasso Regression
7. ElasticNet Regression

We also propose to implement the hybrid Lasso and Gradient boosting regression model to predict individual house price as described in the paper : A Hybrid Regression Technique for House Prices Prediction.
Refer : https://ieeexplore-ieee-org.proxy.lib.csus.edu/stamp/stamp.jsp?tp=&arnumber=8289904
Feature analysis : PCA & Univariate ANOVA f-test.
Boosting algorithms : XGBoost, LightGBM, Cat Boost.
Finally, we would be comparing the performances using the Evaluation plan described below.

Evaluation plan: In order to evaluate that how well our model is doing we will focus on below performance metrics: RMSE score.

Task Division:

Tushar Gandhi :

1. Analysis of The Ames Housing dataset and figuring out the attributes need to be considered and finalize the approach of how to solve the problem.
2. Analysis of The Boston Housing dataset and figuring out the attributes need to be considered and finalize the approach of how to solve the problem.
3. Data cleaning of The Ames Housing dataset.
4. Implementation of the 7 traditional regression models stated above on The Ames Housing dataset.
5. Research and planning of implementation of the hybrid regression technique.
6. Comparison of the performance of all the models.

Akshada Rasam :

1. Analysis of Singapore Housing Prices and figuring out the attributes need to be considered and finalize the approach of how to solve the problem.
2. Data cleaning of Singapore Housing Prices dataset and The Boston Housing dataset.
3. Implementation of the 7 traditional regression models stated above on the Singapore Housing Prices dataset and The Boston Housing dataset.
4. Implementation of the hybrid regression technique.
5. Comparison of the performance of all the models.