

Project Title: “House Prices: Advanced Regression Techniques”

Group Members:

Syed Fahim Shahariar	ID: 2015-3-60-031
Towhiduzzaman	ID: 2016-1-60-031
Md. Khalid Hassan	ID: 2016-2-60-072

Project Description: Nowadays buying houses has to be one of the biggest expenditure one makes in their lifetime. Often time people chose agents to show them their desired type of house within the budget. As the industry is blooming, oftentimes people are misled because of a higher commission pay. In this project we will be analyzing Housing prices by their features and attributes so that you get an estimate of what the real rice of the house maybe compared to what the realtor might be offering you using some of the basic regression algorithms like Linear Regression, Random Forest, and XG Boost and compare the results.

Dataset: The Kaggle dataset that is we intend to use in this project is mainly composed of multiple categorical variables stored as integers and factors, both discrete and continuous. The data set which will be used in this project describes the sale of individual residential property in Ames, Iowa from 2006 to 2010. The original training dataset has a total of one thousand one hundred and sixty observations with eighty one variables.

Methods:

A. Linear Regression

Linear Regression is an ML algorithm based on supervised learning. We will apply it in this project. Regression model generates a forecast value based on independent variables. It is mostly used for finding out the relationship between variables and predicting. It regression performs the work to predict an assigned variable value based on a given independent variable. So, this regression technique finds out a linear relationship between input and output.

B. XG Boost

XG Boost is a decision-tree based resembling ML algorithm that uses a gradient boosting framework. Being a non-parametric learning model, XG Boosting does not require variables to follow a normal distribution line. And this is crucial since housing prices cannot be explained simply through a straight line.

C. Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. We will use it for prediction.

D. Lasso & ridge regression

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models. Ridge Regression is a technique for analyzing multiple regression data that suffer from multi-co-linearity.

Tools: PyCharm

Expected Outcome: In this project, we expect to briefly present dataset gathered from Kaggle. We will investigate the properties and the statistics of the dataset and will try to implement some regression algorithms. Overall our expectation is that, if machine can analyze the housing prices prediction, people can minimize the use of realtors and get more accurate house settings in their preferred budget.