# Advanced Regression Techniques : Housing Prices

Ankit Saxena[1]*, Samvat Rastogi[2],Shailendra Patil[3]

**Abstract**
This project deals with predicting the Sales Price of individual residential properties in Ames, Iowa. The data set consists of both 'Training Data' and Test Data'. The goal is to fit an appropriate regression model using the 'Training Data' and then predict the Sale Price for the given 'Test Data' based on the model. The 'Training Data' consist of 1460 observation and 81 Attributes including the Sales Price and 'Test Data' contains 1459 observations with 80 attributes which does not include Sales Price. Before applying any model, the data is preprocessed to take care of missing values and also to reduce number of attributes being used and then various regression methods like 'Linear', 'Random Forest' 'Gradient Boosting' , 'Xtreme Gradient Boosting' is applied and compared, and the model with least error rate is then used to predict the Sales Price for 'Test Data'.

**Keywords**
Keyword1 — Keyword2 — Keyword3

[1] *Data Science Science, School of Informatics and Computing, Indiana University, Bloomington, IN, USA*
[2] *Data Science Science, School of Informatics and Computing, Indiana University, Bloomington, IN, USA*
[3] *Data Science Science, School of Informatics and Computing, Indiana University, Bloomington, IN, USA*
*__Corresponding author__: john@smith.com

## Contents



## Introduction

"The Housing Market refers to the supply and demand for houses, usually in a particular country or region. A key element of the housing market is the average house prices and trend in house prices"[1]. For many people buying a property is one of the important decisions in life. There are many factors, like Area, Neighbourhood, Connectivity, Transportation ,House Price etc that impacts this decision. Hence it becomes very important to know the trend in which the houses are being sold, so that as a buyer we might have a rough idea what can be the expected Price of a house. The need for predictive analysis is what is required here. As a analyst , we build various model and predict the rough estimate of the prices, which then is used by the property evaluators for their clients.

## 1. Background

### 1.1 Data

The data that has been provided for analysis consists of 3 files

1. **train.csv**
2. **test.csv**
3. **data_description.txt**

Train and Test data files consists of a column named ID, which can be used to uniquely identify every house and 79 attributes that can be used to predict the Sale Price of the house. Train data file consists of an additional column named Sale Price. The Sale Price of the houses in the training data file can be used to build the model for prediction of the Sale Price of the Test data. The data_description file provides all necessary information regarding the attributes. The data was downloaded from kaggle.

### 1.2 Data Description

The training data has 79 attributes excluding the Sales Price and a total of 1460 values. The test data has 78 attributes and a total of 1459 values. Data consists of **43 categorical attributes** and **36 numerical variables**.
The **categorical variables** are: MSZoning, Street, Alley, LotShape, LandContour, Utilities, LotConfig, LandSlope, Neighborhood, Condition1, Condition2, BldgType, HouseStyle, RoofStyle, RoofMatl, Exterior1st, Exterior2nd, MasVnrType, ExterQual, ExterCond, Foundation, BsmtQual, BsmtCond,

BsmtExposure, BsmtFinType1, BsmtFinType2, Heating, HeatingQC, CentralAir, Electrical, KitchenQual, Functional, FireplaceQu, GarageType, GarageFinish, GarageQual, GarageCond, PavedDrive, PoolQC, Fence, MiscFeature, SaleType SaleCondition
The **numerical attributes** are: MSSubClass, LotFrontage, LotArea, OverallQual, OverallCond, YearBuilt, YearRemodAdd, MasVnrArea, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF, X1stFlrSF, X2ndFlrSF, LowQualFinSF, GrLivArea, BsmtFullBath, BsmtHalfBath, FullBath, HalfBath, BedroomAbvGr, KitchenAbvGr, TotRmsAbvGrd, Fireplaces, GarageYrBlt, GarageCars, GarageArea, WoodDeckSF, OpenPorchSF, EnclosedPorch, X3SsnPorch, ScreenPorch, PoolArea, MiscVal, MoSold, YrSold

## 1.3 Data Preprocessing

The biggest task in any Data Mining problem is 'Data Preprocessing' . Before we apply any algorithms , we first need to clean the data. The data might contain missing values, outlier etc. There will be different kinds of attribute like continuous variables, categorical variables and so on. Such variables need to be taken care of based on the model we want to apply. Data Preprocessing not only deals with these thing but also feature selection, data reduction etc. Given a set of attribute , not every attribute is important for data analysis or prediction, few attributes might be of no use and few attributes will be very vital for our analysis, so for this purpose we do Feature extraction.

### 1.3.1 Missing Data Imputation

There are many attributes which have NA as one of the accepted value in their domain. These NA's are not missing value but rather explain the absence of a particular feature in the house. The following attributes have NA in their domain: Alley, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2, FireplaceQu, GarageType, GarageFinish, GarageQual, GarageCond, PoolQC, Fence, MiscFeature. The NA values for these attributes were replaced with **NC** while cleaning the data, to differentiate the actual missing values for other attributes with these 14 attributes.After replacing the NA values with NC for these attributes, we found the number of actual NA values or missing values: A total of **357** values are missing in the Training data. A total of **358** values are missing in the Test data. The percentage of missing values in the training data is **0.3018%** The percentage of missing values in the test data is **0.3067%**

The majority of values are missing for the attribute **LotFrontage** The percentage of missing values for LotFrontage in training data is **17.739%** The percentage of missing values for LotFrontage in test data is **15.558%** The mean of the values for attribute **LotFrontage** in training data is **70.05** and the median is **69.00**. The mean of the values for attribute **LotFrontage** in test data is **68.58** and the median is **67.00**. The missing values have been replaced with the mean of the existing values for LotFrontage, as there is not a huge difference between the mean and the median.

The attribute **GarageYrBlt** has the second highest number of missing values. The percentage of missing values for **GarageYrBlt** in training data is **5.547%** The percentage of missing values for **GarageYrBlt** in test data is **5.346%** The mean of the values for attribute **GarageYrBlt** in training data is **1979** and the median is **1980**. The mean of the values for attribute **GarageYrBlt** in test data is **1978** and the median is **1979**. The missing values have been replaced with the mean of the existing values for GarageYrBlt, as there is not a huge difference between the mean and the median. Similarly for all other numerical attributes missing values was replaced by Mean or Median based on data and for categorical attributes the missing values were replaced by the mode of that particular attribute.

### 1.3.2 Feature Selection/Feature Selection

**Feature extraction** starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. Feature extraction is related to dimensionality reduction [5]. When the input data to an algorithm is too large to be processed and it is suspected to be redundant, then it can be transformed into a reduced set of features (also named a features vector). This process is called **feature selection** [5].

One of the methods for data reduction is **Correlation**. Correlation is a statistical technique that can show whether and how strongly pairs of variables are related.
We are using **Pearsons Coeffiecient** and this can be done in **R** using **cor** function. Below is the graph that shows the correlation among all numerical attributes. As seen in the graph,

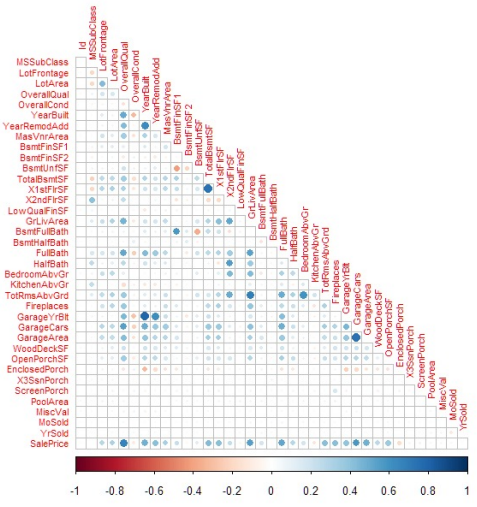


**Figure 1.**
Correlation Between Variables

the attributes OverallQual,GrLiveArea,FullBath, GarageCars are among few attributes which have greater correlation with SalePrice. But the Pearson's cofficient needs linear data and data should be a normal, and as data size is small just by plotting the graphs its not easy to predict if the data is normal and attributes have linear relationship or not, hence correlation isn't a feasible idea here.

So the method we have used for feature extraction is using **Boruta Algorithm**. The algorithm is designed as a wrapper around a Random Forest classication algorithm. It iteratively removes the features which are proved by a statistical test to be less relevant than random probes. The **Boruta package** in **R** provides a convenient interface to the algorithm.

Below is the step wise working of **boruta algorithm**[6]:

1. Firstly, it adds randomness to the given data set by creating shuffled copies of all features (which are called shadow features).
2. Then, it trains a random forest classifier on the extended data set and applies a feature importance measure (the default is Mean Decrease Accuracy) to evaluate the importance of each feature where higher means more important.
3. At every iteration, it checks whether a real feature has a higher importance than the best of its shadow features (i.e. whether the feature has a higher Z score than the maximum Z score of its shadow features) and constantly removes features which are deemed highly unimportant.
4. Finally, the algorithm stops either when all features gets confirmed or rejected or it reaches a specified limit of random forest runs.

The Boruta package was used after the missing values were replaced, as Boruta does not work if we have missing values in data. After running the Boruta function and plotting the graph we get a list of **confirmed**, **tentative** and **rejected** attribute. Below is the graph of these variables

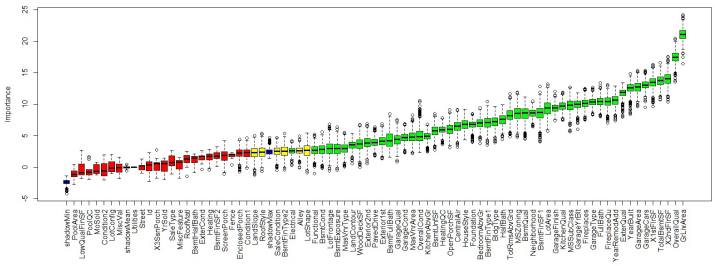

**Figure 2.**
Boruta Plot

The **green** box plots are for the attributes that have been confirmed, the **yellow** for tentative and **red** for rejected attributes.

## 2. Algorithm and Methodology

The project mainly deals with using various Regression techniques like **"Linear"**,**"Random Forest Regression"**,**"Gradient Boosting"**,**"Extreme Gradient Boosting"** for predictive analysis.

**What is Regression Analysis?**

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor). This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables [2].

### 2.1 Regression Techniques

1. **Linear Regression**
   A linear regression establishes a relation between a dependent variable (Y) and one or more independent variable(X) using a best fit straight line,(also known as "Regression Line").
   It is represented by below equation

$$y = a + bx \tag{1a}$$

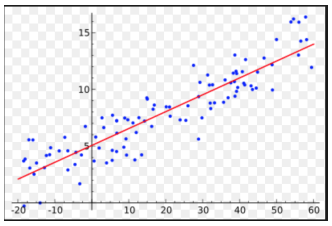


**Figure 3.**
Linear Regression Line

2. **Random Forest Regression**
   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 mean prediction (regression) of the individual trees [3].

3. **Gradient Boosting**
   Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function [4].

4. **LASSO Regression**: lasso (least absolute shrinkage and selection operator) (also Lasso or LASSO) is a regression analysis method that performs both variable

selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces

#### 2.1.1 Subsubsection

Etiam euismod. Fusce facilisis lacinia dui. Suspendisse potenti. In mi erat, cursus id, nonummy sed, ullamcorper eget, sapien. Praesent pretium, magna in eleifend egestas, pede pede pretium lorem, quis consectetuer tortor sapien facilisis magna. Mauris quis magna varius nulla scelerisque imperdiet. Aliquam non quam. Aliquam porttitor quam a lacus. Praesent vel arcu ut tortor cursus volutpat. In vitae pede quis diam bibendum placerat. Fusce elementum convallis neque. Sed dolor orci, scelerisque ac, dapibus nec, ultricies ut, mi. Duis nec dui quis leo sagittis commodo.

**Word** Definition

**Concept** Explanation

**Idea** Text

#### 2.1.2 Subsubsection

Aliquam lectus. Vivamus leo. Quisque ornare tellus ullamcorper nulla. Mauris porttitor pharetra tortor. Sed fringilla justo sed mauris. Mauris tellus. Sed non leo. Nullam elementum, magna in cursus sodales, augue est scelerisque sapien, venenatis congue nulla arcu et pede. Ut suscipit enim vel sapien. Donec congue. Maecenas urna mi, suscipit in, placerat ut, vestibulum ut, massa. Fusce ultrices nulla et nisl.

- First item in a list
- Second item in a list
- Third item in a list

#### 2.1.3 Subsubsection

Etiam ac leo a risus tristique nonummy. Donec dignissim tincidunt nulla. Vestibulum rhoncus molestie odio. Sed lobortis, justo et pretium lobortis, mauris turpis condimentum augue, nec ultricies nibh arcu pretium enim. Nunc purus neque, placerat id, imperdiet sed, pellentesque nec, nisl. Vestibulum imperdiet neque non sem accumsan laoreet. In hac habitasse platea dictumst. Etiam condimentum facilisis libero. Suspendisse in elit quis nisl aliquam dapibus. Pellentesque auctor sapien. Sed egestas sapien nec lectus. Pellentesque vel dui vel neque bibendum viverra. Aliquam porttitor nisl nec pede. Proin mattis libero vel turpis. Donec rutrum mauris et libero. Proin euismod porta felis. Nam lobortis, metus quis elementum commodo, nunc lectus elementum mauris, eget vulputate ligula tellus eu neque. Vivamus eu dolor.

### 2.2 Subsection

Nulla in ipsum. Praesent eros nulla, congue vitae, euismod ut, commodo a, wisi. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Aenean nonummy magna non leo. Sed felis erat, ullamcorper in, dictum non, ultricies ut, lectus. Proin vel arcu a odio lobortis euismod. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Proin ut est. Aliquam odio. Pellentesque massa turpis, cursus eu, euismod nec, tempor congue, nulla. Duis viverra gravida mauris. Cras tincidunt. Curabitur eros ligula, varius ut, pulvinar in, cursus faucibus, augue.

Nulla mattis luctus nulla. Duis commodo velit at leo. Aliquam vulputate magna et leo. Nam vestibulum ullamcorper leo. Vestibulum condimentum rutrum mauris. Donec id mauris. Morbi molestie justo et pede. Vivamus eget turpis sed nisl cursus tempor. Curabitur mollis sapien condimentum nunc. In wisi nisl, malesuada at, dignissim sit amet, lobortis in, odio. Aenean consequat arcu a ante. Pellentesque porta elit sit amet orci. Etiam at turpis nec elit ultricies imperdiet. Nulla facilisi. In hac habitasse platea dictumst. Suspendisse viverra aliquam risus. Nullam pede justo, molestie nonummy, scelerisque eu, facilisis vel, arcu.

Curabitur tellus magna, porttitor a, commodo a, commodo in, tortor. Donec interdum. Praesent scelerisque. Maecenas posuere sodales odio. Vivamus metus lacus, varius quis, imperdiet quis, rhoncus a, turpis. Etiam ligula arcu, elementum a, venenatis quis, sollicitudin sed, metus. Donec nunc pede, tincidunt in, venenatis vitae, faucibus vel, nibh. Pellentesque wisi. Nullam malesuada. Morbi ut tellus ut pede tincidunt porta. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam congue neque id dolor.

Donec et nisl at wisi luctus bibendum. Nam interdum tellus ac libero. Sed sem justo, laoreet vitae, fringilla at, adipiscing ut, nibh. Maecenas non sem quis tortor eleifend fermentum. Etiam id tortor ac mauris porta vulputate. Integer porta neque vitae massa. Maecenas tempus libero a libero posuere dictum. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Aenean quis mauris sed elit commodo placerat. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Vivamus rhoncus tincidunt libero. Etiam elementum pretium justo. Vivamus est. Morbi a tellus eget pede tristique commodo. Nulla nisl. Vestibulum sed nisl eu sapien cursus rutrum.

Nulla non mauris vitae wisi posuere convallis. Sed eu nulla nec eros scelerisque pharetra. Nullam varius. Etiam dignissim elementum metus. Vestibulum faucibus, metus sit amet mattis rhoncus, sapien dui laoreet odio, nec ultricies nibh augue a enim. Fusce in ligula. Quisque at magna et nulla commodo consequat. Proin accumsan imperdiet sem. Nunc porta. Donec feugiat mi at justo. Phasellus facilisis ipsum quis ante. In ac elit eget ipsum pharetra faucibus. Maecenas viverra nulla in massa.

Nulla ac nisl. Nullam urna nulla, ullamcorper in, interdum sit amet, gravida ut, risus. Aenean ac enim. In luctus. Phasellus eu quam vitae turpis viverra pellentesque. Duis feugiat felis ut enim. Phasellus pharetra, sem id porttitor sodales, magna nunc aliquet nibh, nec blandit nisl mauris at pede. Suspendisse risus risus, lobortis eget, semper at, imperdiet sit amet, quam. Quisque scelerisque dapibus nibh. Nam enim. Lorem ipsum dolor sit amet, consectetuer adipisc-

ing elit. Nunc ut metus. Ut metus justo, auctor at, ultrices eu, sagittis ut, purus. Aliquam aliquam.

Etiam pede massa, dapibus vitae, rhoncus in, placerat posuere, odio. Vestibulum luctus commodo lacus. Morbi lacus dui, tempor sed, euismod eget, condimentum at, tortor. Phasellus aliquet odio ac lacus tempor faucibus. Praesent sed sem. Praesent iaculis. Cras rhoncus tellus sed justo ullamcorper sagittis. Donec quis orci. Sed ut tortor quis tellus euismod tincidunt. Suspendisse congue nisl eu elit. Aliquam tortor diam, tempus id, tristique eget, sodales vel, nulla. Praesent tellus mi, condimentum sed, viverra at, consectetuer quis, lectus. In auctor vehicula orci. Sed pede sapien, euismod in, suscipit in, pharetra placerat, metus. Vivamus commodo dui non odio. Donec et felis.

Etiam suscipit aliquam arcu. Aliquam sit amet est ac purus bibendum congue. Sed in eros. Morbi non orci. Pellentesque mattis lacinia elit. Fusce molestie velit in ligula. Nullam et orci vitae nibh vulputate auctor. Aliquam eget purus. Nulla auctor wisi sed ipsum. Morbi porttitor tellus ac enim. Fusce ornare. Proin ipsum enim, tincidunt in, ornare venenatis, molestie a, augue. Donec vel pede in lacus sagittis porta. Sed hendrerit ipsum quis nisl. Suspendisse quis massa ac nibh pretium cursus. Sed sodales. Nam eu neque quis pede dignissim ornare. Maecenas eu purus ac urna tincidunt congue.

Donec et nisl id sapien blandit mattis. Aenean dictum odio sit amet risus. Morbi purus. Nulla a est sit amet purus venenatis iaculis. Vivamus viverra purus vel magna. Donec in justo sed odio malesuada dapibus. Nunc ultrices aliquam nunc. Vivamus facilisis pellentesque velit. Nulla nunc velit, vulputate dapibus, vulputate id, mattis ac, justo. Nam mattis elit dapibus purus. Quisque enim risus, congue non, elementum ut, mattis quis, sem. Quisque elit.

## 3. Experiments and Results

## 4. Summary and Conclusions

## Acknowledgments

So long and thanks for all the fish [**?**].

[1]http://www.economicshelp.org/blog/glossary/definition-of-the-housing-market/
[2]https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression
[3]https://en.wikipedia.org/wiki/Random_Forest
[4]https://en.wikipedia.org/wiki/Gradient_boosting
[5]https://en.wikipedia.org/wiki/Feature_extraction
[6]https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/