Name: Nthimbo Gift Tembo

ID: 1738522

Kaggle ID: CCode

Kaggle Score: 0.12540

Title: Kaggle House Price Prediction

Abstract and Introduction

The whole purpose of the project was to predict Kaggle House Prices based on multiple features using various regression methodologies while at the same time considering efficiency of the algorithm and good yield of results.

Methodology

This project was done in series of countable steps as depicted below,

Step1: Selection of Data

In the first step I checked if all the features I was going to use contained missing values and the kind of data type they hold (Numbers or categorical). To deal with the missing values I replaced the missing values with the most frequent values. On the side of categorical values, I converted all the categorical values to numerical values as I decided to use all the features while at the same time being cautious that the score might be lowered if a feature is not relevant. I reframed all the categorical values so that I can get a better structure of the data when using the model as it would only use numerical values.

Step2: Pre-processing by Standardization

After choosing the features to use in the model, I had to check the data if it contained outliers using box plots as are depicted in the graphs. From the plots, I observed from the initial data that many of the features mostly contained some outliers. I removed the outliers by removing all data that deviates from the Mean by a larger number than what I specified for that specific data. I used different values as standard deviation for removing outliers as I observed that using the same value for the deviation gives different results that did not give the best prediction. After doing all the necessary cleaning of the training data, I divided the data into training and test set, where 80% was for training and 20% for testing purposes.

Step 3: Transform of the Data

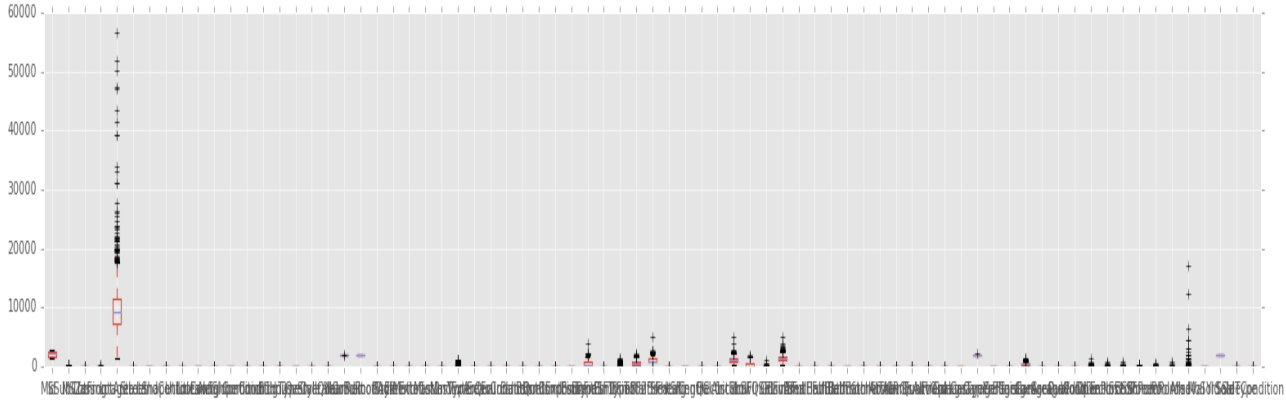
After pre-processing of the data as described above I transformed the data to have better scale using different methods. I considered this step since by transforming variables I would also remove the potential outliers which didn't get removed. I used natural Log to reduce variation which might be caused by extreme values. I also did imputation of the missing values to remove outliers by replacing them with the most frequent values.

Step 4: Data Modelling

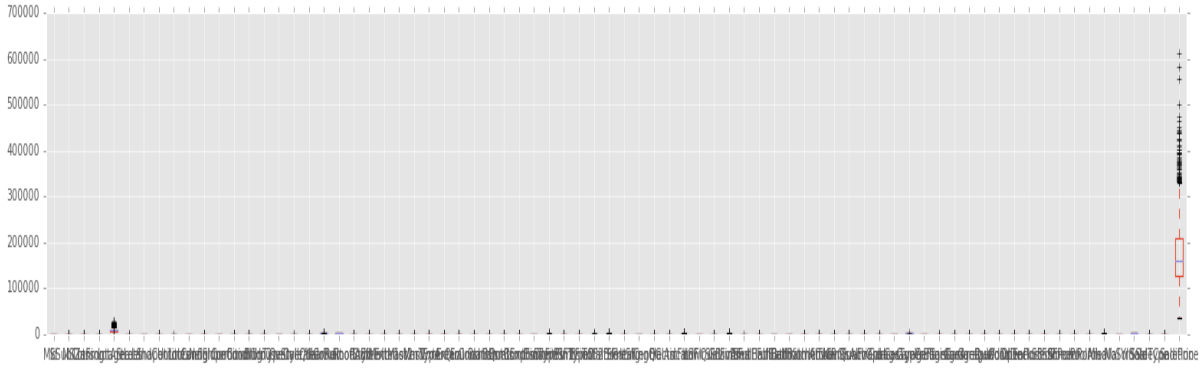
This is the final step where I initialised the model to use. I did consider different regression methods such like, Lasso, Linear Regression, Ridge etc. However, I realised that different models give different results according to how the data has been prepared. For my data Ridge and Linear Regression were giving me the acceptable performance unlike Lasso. Therefore, overall in the project I used Ridge regression for prediction. The detailed graphical results have been included in the next page

Graphs of the Project

1. Graph of the un processed data with outliers using box plot. With



2. Plot after removing outlier



3. Graph with predictions of House prices