House Prices: Advanced Regression Techniques

Aditya Gouroju

October 29 2020

1 Introduction

1.1 Problem

Prediction of house prices with 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa

1.2 Interest

People who are moving to residential homes in Ames, Iowa would be very much interested in finding out the home which is best suitable to them and approximately how much it costs

2 Data Acquisition and Preprocessing

2.1 Data sources

Both Training and Test Datasets are provided in the competition in ".csv" format.An Additional Data Description text file is also provided for the Datasets.

2.2 Data Cleaning

The provided dataset is converted into to Dataframe using the pandas library. The DataFrame consisted of 81 columns in which two of the columns are "Id" and the target feature "SalePrice" which are dropped

Both the training and Test datasets are combined as it is more effective to preprocess them at a time rather than preprocessing them individually. Columns with high "Null" data ratio and which are irrelevant to the target are dropped and other columns with low "Null" data are filled with that respective columns mode/median depending upon the type of feature it is and the description.

In initial attempts I created new columns for NaN entries assuming they were important but it didn't work effectively

2.3 Feature Selection

The columns with object datatype are **One Hot Encoded** which are later used as the features for the model.The logarithmic values of the target feature shows more normal distribution than the normal values.

All the columns which are One Hot Encoded and the remaining features are used for predicting the logarithmic values of the house SalePrice which is now the target feature for the model

3 Methodology

3.1 Advanced Regression

Regression machine learning model for this problem, The main aim is to predict the prices of houses.Sci-kit learn library is used for the regression

Stacked Regression with multiple models like Gradient Boosters ,Ridge, Regressions and tuning the hyper parameters of the models for improved performance

In Initial attempts i used basic single linear models like Ridge and Lasso which did not show much effect.

The evaluation is based on **Root Mean Squared Error**.

4 Results & Conclusion

By this model I was able to score 0.12023 in the kaggle competition (Username:adityagouroju)