Sathya
16 January 2020

AMES Housing Price Prediction

Problem Statement

The second Data Science Project aims to develop the algorithm to estimate the price of residential houses based on fixed features i.e. characteristics that cannot be easily renovated (e.g. location, square feet, number of bedrooms and bathrooms).This project uses the Ames housing data available on Kaggle, which includes 81 features describing a wide range of characteristics of 1,460 homes in Ames, Iowa sold between 2006 and 2010.

What features of a house are most important (have the highest correlation) in predicting its price in the Ames market?

Exploratory Data Analysis:

The histogram of the sales price of the plot is as below.

The following assumptions are made based on the EDA:

Validated the data by profiling and expolatory Data Analysis, and have the following outcomes.

1.Utilities feature can be eliminated , as most of column has value AllPub.
2.Street feature can be eliminated - as most of values has Pave
3.Neighborhood is included in the feature selection whereas the "Condition 1" and "Condition 2" features realted to Neighborhood are elimated.
4.Year Build, Year Modified, Year Sold are computed to "Year Since Modified" and "Years to Sell"
"Year Since Modified" = "Year Built" - "Years Remod"
"Years to Sell" = "Year Sold" - "Years Built"
5."Roof Matl" feature can be eliminated - as most of values has CompShg
6."Mas Vnr Type" is omitted as we consider "Mas Vnr Area"
7.All this features are eliminated as we consider the feature Total Bsmt Area.
1)Bsmt Exposure
2)BsmtFin Type 1
3)BsmtFin SF 1
4)BsmtFin Type 2
5)BsmtFin SF 2
8.As we considered the feature "overall SF" we elimibnate the following columns
i)1st Flr SF
ii)2nd Flr SF
9.We considered the "Garage Area" and eliminate the other columns.
1)Garage Type
2)Garage Yr Blt
3)Garage Finish
4)Garage Cars
10.All the other miscillneous columns are omitted.
1)Pool QC
2)Fence
3)Misc Feature
4)Misc Val
5)Mo Sold

Data Cleaning:

The following cleaning process have been made to the dataset.

1. Years to sell & Years to Remod are calculated based on Year Built , Years Remod , Years Sold
2. For all the categorical Columns which are not ordinal columns we created the dummies for each vlaue of the column. The columns which falls under this category are as follows: Lot Shape, Land Contour, Lot Config, Land Slope, MS Zoning, House Style, Roof Style

3. For all the categorical column which are ordinal number which denotes the rating are converted to number by assigning th values.The following columns were converrted to the ordinal number columns. Exter Qual, Exterr Cond, Bsmt Qual, Bsmt Cond , Heating QC,Kitchen Qual,Fireplace Qu,Garage Qual

4. Added Zero the below columns for NAN 'Mas Vnr Area' , Lot Frontage

5. Add Column Porch Area adding the sum of Open Porch SF,Enclosed Porch,3Ssn Porch ,Screen Porch

Model Evaluation:

Model1 : Linear Regression Model with 94 Attributes. And RMSE : 3476171755886500.0

Model2 : Lasso Regression Model with 94 Attributes. And RMSE : 27780.05705795639

Based On Model2 , The Lasso gives the Co- efficient as Zero for some Features. Removed the features with zero Co - Efficient

Model 3 : Ridge Regression Model with 27 Attributes. And RMSE : 29007.241584468393

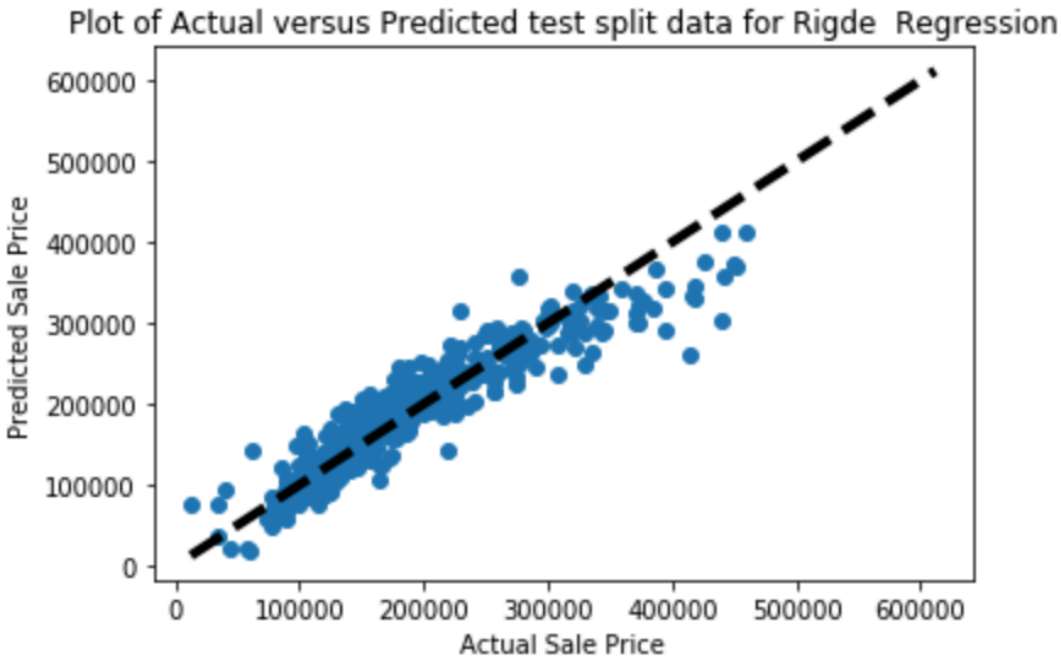
Model4 : Elastic Net Regression Model with 27 features. And RMSE : 29267.20525682117

Based on the Model 4, Removed the features with ElasticNet Co-efficent less than 2300 and greater than -2000

I tried to improve the model performance by grid searching on the ElasticNet and LassoCV and removed some of features . Knowing that Lasso regularisation results in a sparser model and deals with multicollinearity between predictors well, I decided to choose the optimum Ridge model for predicting the sale price of residential houses based on fixed features. This model achieved a very similar R2 score (0.869) to the linear regression model above – but importantly, I removed 6 predictors. The 20 largest model coefficients in terms of absolute value, are displayed in the bar plot below.

Final Model :

The final model is Ridge Regression with 20 features. The below is plot between Actual Sale Price with Predicted Sale price for the test data based on train- test Split Data



Prediction for the Test Dataset & Kaggle Submission

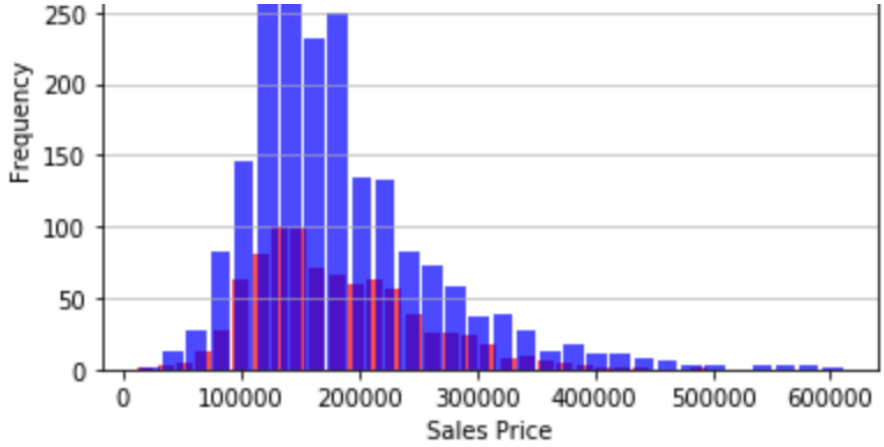
The predictions of the sales Price for the test data were predicted based on the final model and submitted the final csv to the Kaggle competition.

The histogram of the sale price of the train dataset with the predicted test dataset.

Name	Submitted	Wait time	Execution time	Score
test_prediction_final.csv	a few seconds ago	0 seconds	0 seconds	34228.39262

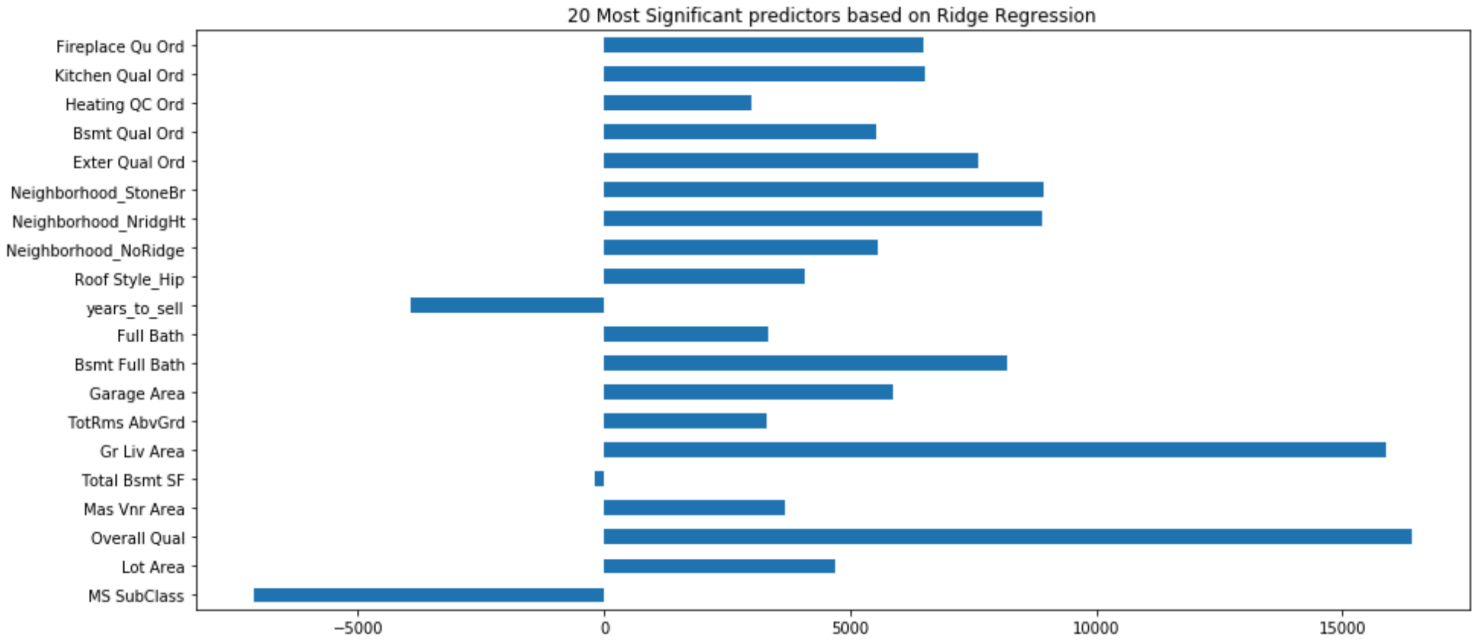
Complete

Jump to your position on the leaderboard



Business Recommendations

- Which features appear to add the most value to a home. **Overall Qual, Gr Liv Area , Neighborhood**



- Which features hurt the value of a home the most? **Years to Sell - Which is Year Build - Year Sold**

- What are things that homeowners could improve in their homes to increase the value? **Overall Quality**

- What neighborhoods seem like they might be a good investment? **StoneBr, NridgHt , NoRidge likely be good investment.**

- Do you feel that this model will generalize to other cities? How could you revise your model to make it more universal OR what date would you need from another city to make a comparable model?**Yes.**