AMES Housing Data Analysis

Description:

In this project, we will be looking at the AMES Housing dataset. The dataset contains a sample of around 1,400 houses, 79 features (variables) and the target column called “Sales Price”. The features are deemed to be the independent criterion which influence the sales price of the house. For instance, from observing the coefficients of the linear regression, we see that the garage quality has the power to substantially reduce the sales price, and the roof material has the power to heavily increase the sales price.

Out of the 79 features, 43 are categorical features that will be dealt with using the pandas get_dummies. For interpretability, a summary of all attributes is not possible as the number of features is so large.

Objective:

The main objectives of the analysis are:

- Produce a regression model that predicts the sales price of a house.
- Compare three different models to find the best model and highlight any possible flaws.
- Use parameter tuning to find the optimal set of hyperparameters.
- Further possible steps.

Analysis steps:

Missing Data

The first step was checking to see if there was missing data. In this instance, there was no missing data so we moved onto the next step.

Encoding Categorical Data

As mentioned previously, there were 43 categorical features. After the encoding, this resulted in a total of 251 features.

Feature Scaling

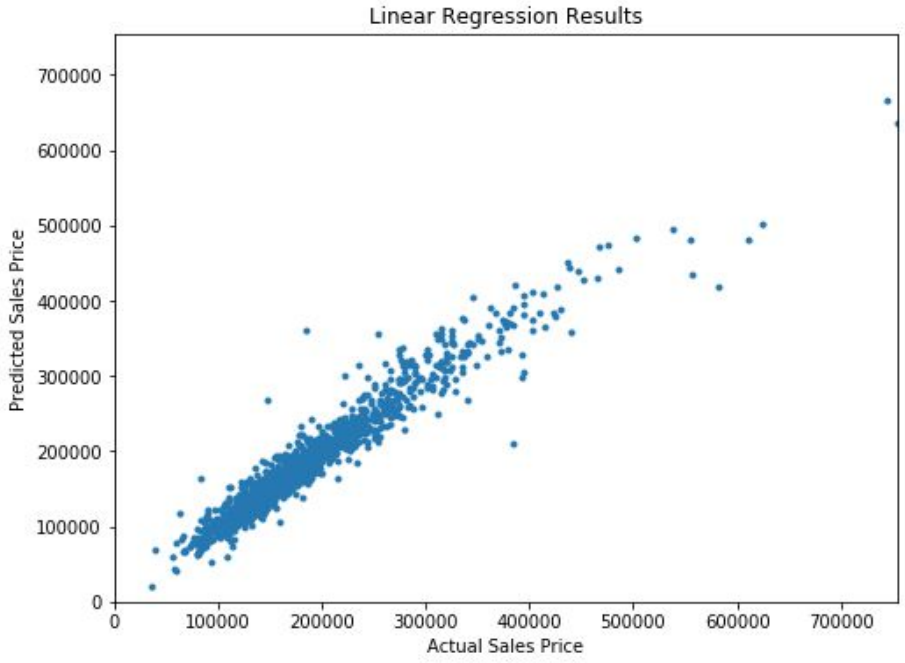
Although not necessary for linear regression, we scaled the values, fit the entire dataset to the model and predicted the dataset. The below shows a short summary of the most impacting coefficients of the linear regression:

147	GarageQual_TA	-24549.449822
144	GarageQual_Fa	-21939.456460
171	KitchenQual_TA	-11864.500818
170	KitchenQual_Gd	-11524.713090
145	GarageQual_Gd	-11149.884887

...	...	...
28	PoolArea	26321.276440
229	RoofMatl_WdShake	33553.116951
230	RoofMatl_WdShngl	41451.642204
228	RoofMatl_Tar&Grv	49062.795491
224	RoofMatl_CompShg	76668.421574

Looking at the strength of the standardized coefficients GarageQual_TA, GarageQual_Fa, RoofMatl_Tar and RoofMatl_CompShg, we deem them the 'most impactful'. Sklearn does not have built in statistical significance of each of these variables which would aid in making this claim stronger/weaker.

Sum of coefficients: 1060631.6918750189
Total number of coefficients: 251
Number of coefficients not equal to 0: 251 (Expected due to no feature selection)
Linear Regression R2 score is: 0.9307954458984128



From the scatter plot above, we observe that the actual sales and predicted sales demonstrate a clear linear relationship with only a slight spread. This is inferred by the high R2 score.

In order to fit LASSO and Ridge regression, we first need to add polynomial terms.

<u>Adding Polynomial Terms</u>

Polynomial terms were added by importing Polynomial Features from sklearn. The degree that was given was 2. After adding polynomial terms, we performed Feature Scaling. ***Feature Scaling is compulsory with LASSO and Ridge.***

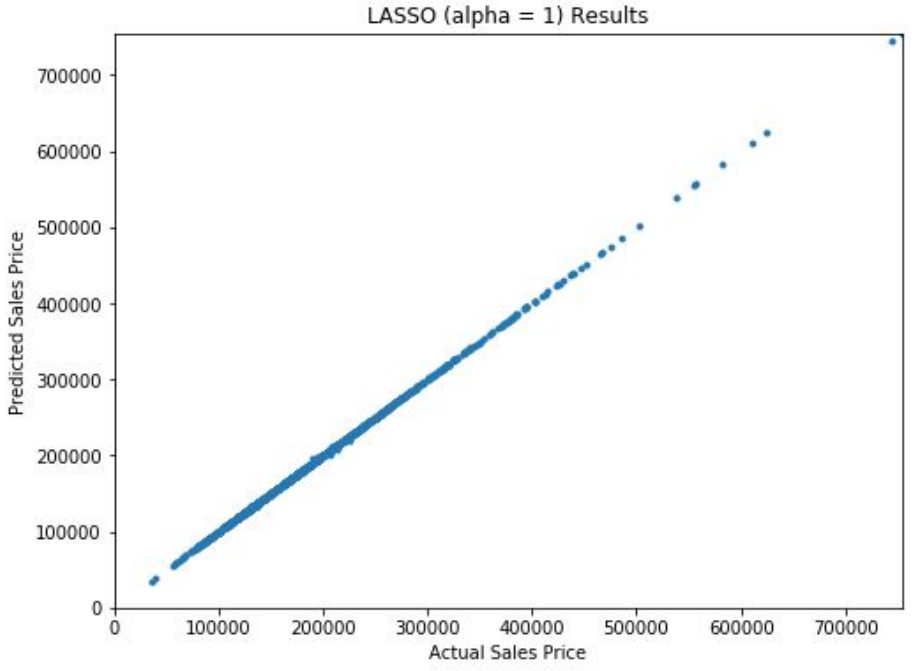
LASSO (alpha = 1)

17	KitchenAbvGr	-10853.557275
15	GrLivArea	-7902.338000
3	BedroomAbvGr	-7450.047762
9	EnclosedPorch	-5873.875496
222	PoolQC_Gd	-4880.172726
...	...	...
12	GarageArea	7991.038279
8	BsmtUnfSF	11203.123351
4	BsmtFinSF1	12186.380613
1	2ndFlrSF	34465.765232
0	1stFlrSF	44808.255274

Looking at the strength of the standardized coefficients KitchenAbvGr, GrLivArea, 2ndFlrSF and 1stFlrSF, we deem them the 'most impactful'. We also observe that the coefficients have a smaller range in comparison to the linear regression.

Sum of coefficients: 1853098.2120321642
Total number of coefficients: 31877
Number of coefficients not equal to 0: 5198
Lasso R2 for alpha = 1 is: 0.999977381920978

From the scatter plot below, we observe that the actual sales and predicted sales demonstrate a clear linear relationship. This is inferred by the extremely high R2 score. The R2 score is much higher than the linear regression R2 score, this being visually demonstrated as the spread of the data points in the LASSO regression is infinitesimal in comparison to linear regression spread of the data points.



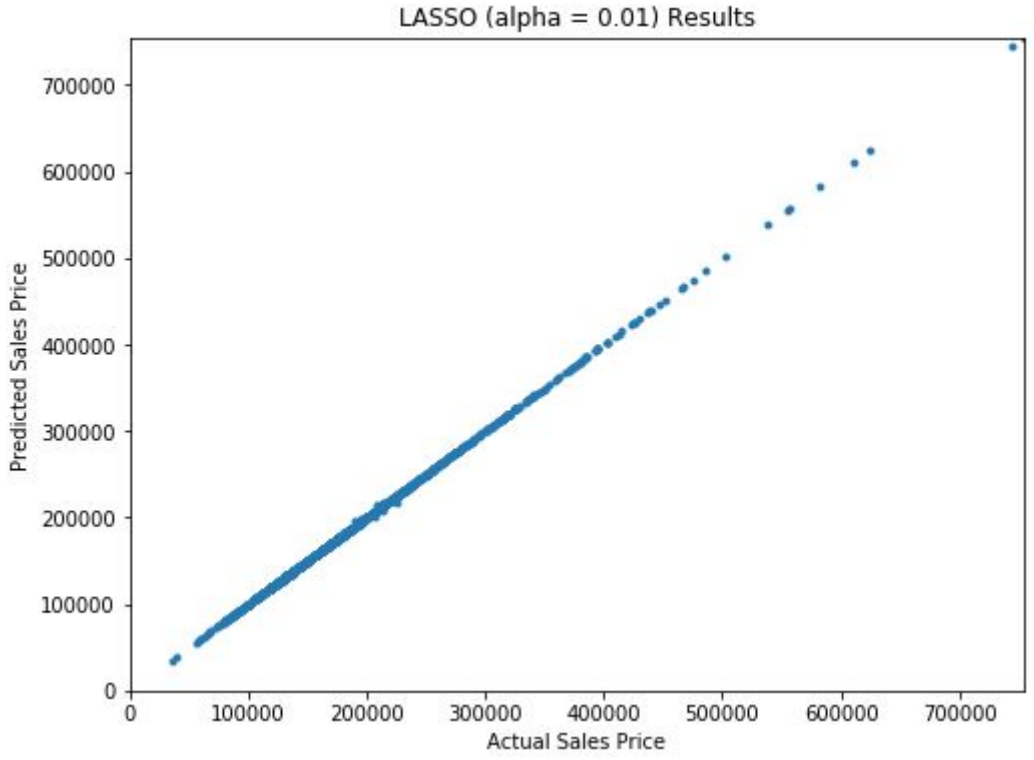
LASSO (alpha = 0.01)

17	KitchenAbvGr	-11671.168388
15	GrLivArea	-10096.826682
3	BedroomAbvGr	-9554.946462
31	TotalBsmtSF	-6980.748401
9	EnclosedPorch	-6956.716021
...	...	...
11	FullBath	8696.505723
8	BsmtUnfSF	11224.211950
4	BsmtFinSF1	12974.595184
1	2ndFlrSF	35814.384035
0	1stFlrSF	47589.344310

We observe that the coefficients are in the same order as LASSO (alpha = 1) however, the coefficients have a larger range/magnitude.

Sum of coefficients: 4142793.33288558
Total number of coefficients: 31877
Number of coefficients not equal to 0: 19980
Lasso R2 for alpha = 0.01 is: 0.9999650095628029

With more regularization (higher alpha) we will expect the penalty for higher weights to be greater and thus the coefficients to be pushed down. Thus a higher alpha means lower magnitude with more coefficients pushed down to 0. This is validated by there being nearly 4 times as many coefficients that are non zero in comparison to LASSO (alpha = 1).



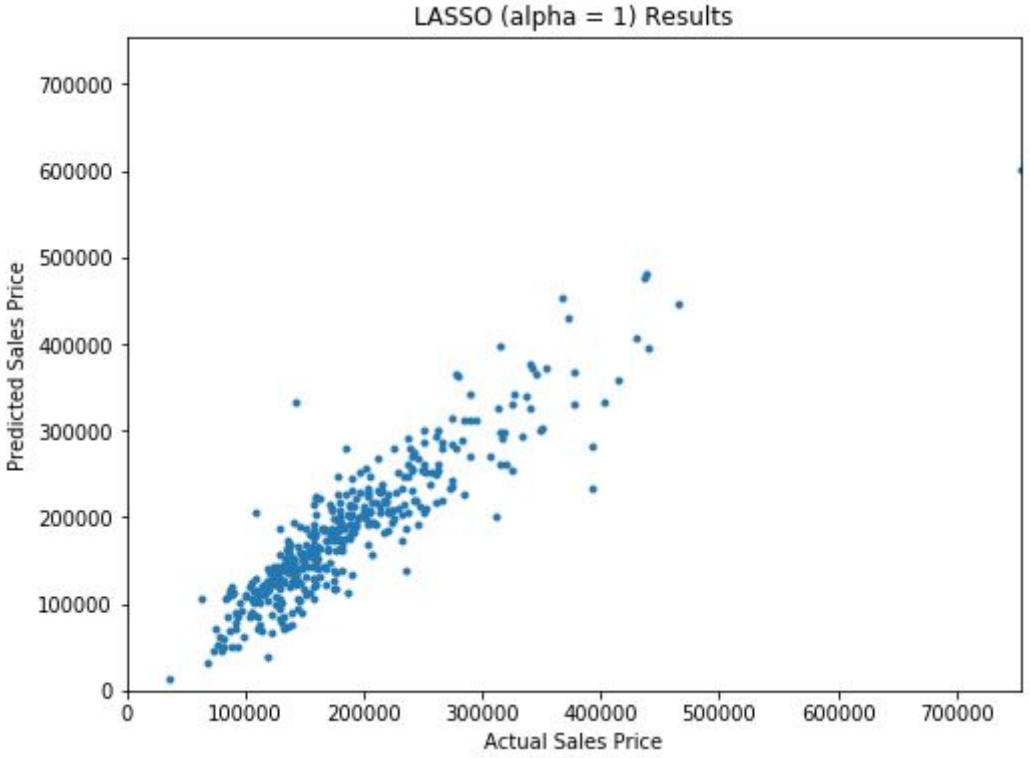
Model Review

Although we have retrieved a great performance for the model, we have not actually tested the model against any new data. Therefore, a flaw in our model is that it is prone to overfitting. We will now split the dataset into the training set and test set and evaluate the models performance.

LASSO (alpha = 1)

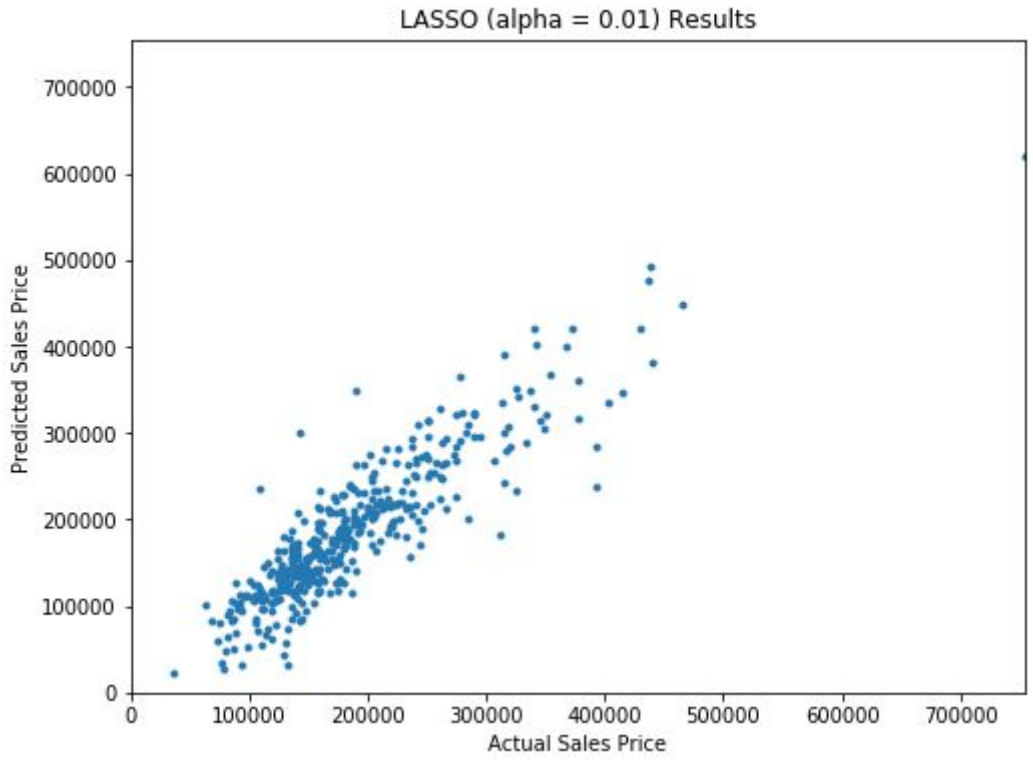
Sum of coefficients: 1510404.3594291694
Total number of coefficients: 31877
Number of coefficients not equal to 0: 4348
Lasso R2 for alpha = 1 is: 0.5334352932323387

We observe that the R2 score has decreased substantially when predicting the test set results. The scatter plot below shows the results, the spread visually depicting the lower R2 score.



LASSO (alpha = 0.01)

Sum of coefficients: 3525350.5082896915
Total number of coefficients: 31877
Number of coefficients not equal to 0: 18728
Lasso R2 for alpha = 0.01 is: 0.38713020386395114



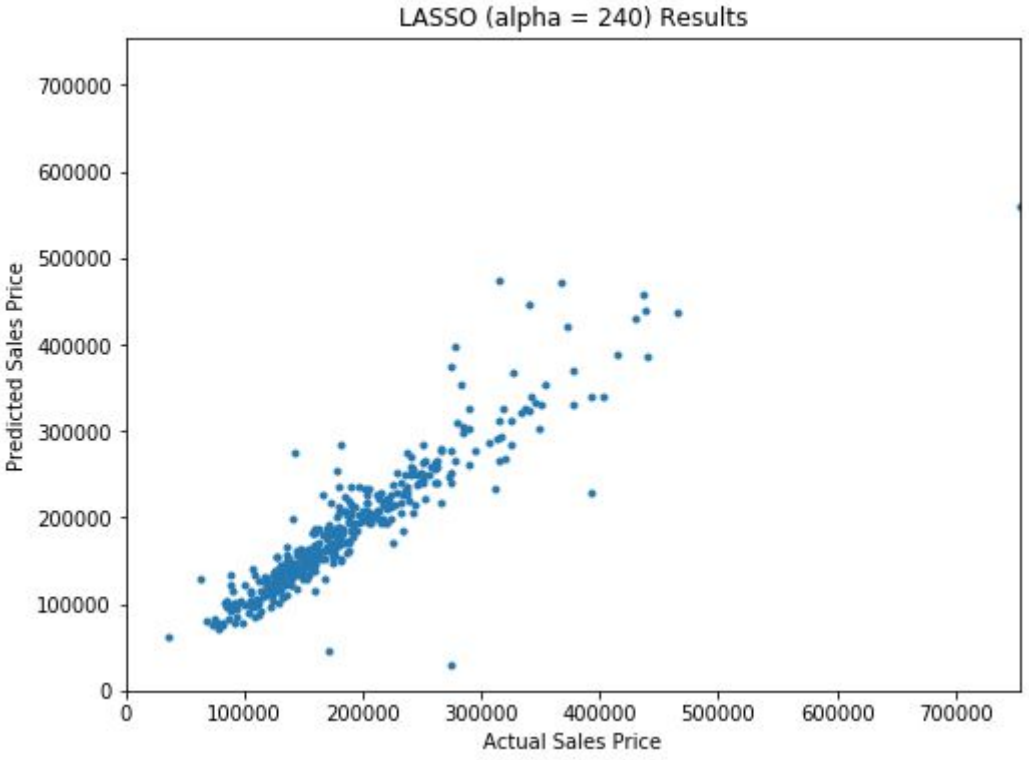
We observe that the R2 score has decreased substantially when predicting the test set results in comparison to LASSO (alpha = 1). As expected, due to the lower regularisation the number of coefficients that are non zero have increased nearly 4 times.

Cross Val Score - Parameter Tuning : LASSO (alpha = 240)

Performing cross validation from importing LassoCV a few times has given us an alpha of 240. The summary obtain is below:

Sum of coefficients: 368406.0279454886
Total number of coefficients: 31877
Number of coefficients not equal to 0: 841
Lasso R2 for alpha = 240 is: 0.844101270644865

Due to the very high alpha value, we observe that the number of non zero coefficients has dropped substantially to 841. In addition, the R2 score has improved dramatically. This can be observed visually from the below scatter plot as most of the points are concentrated on a much straighter line with a reduced spread.

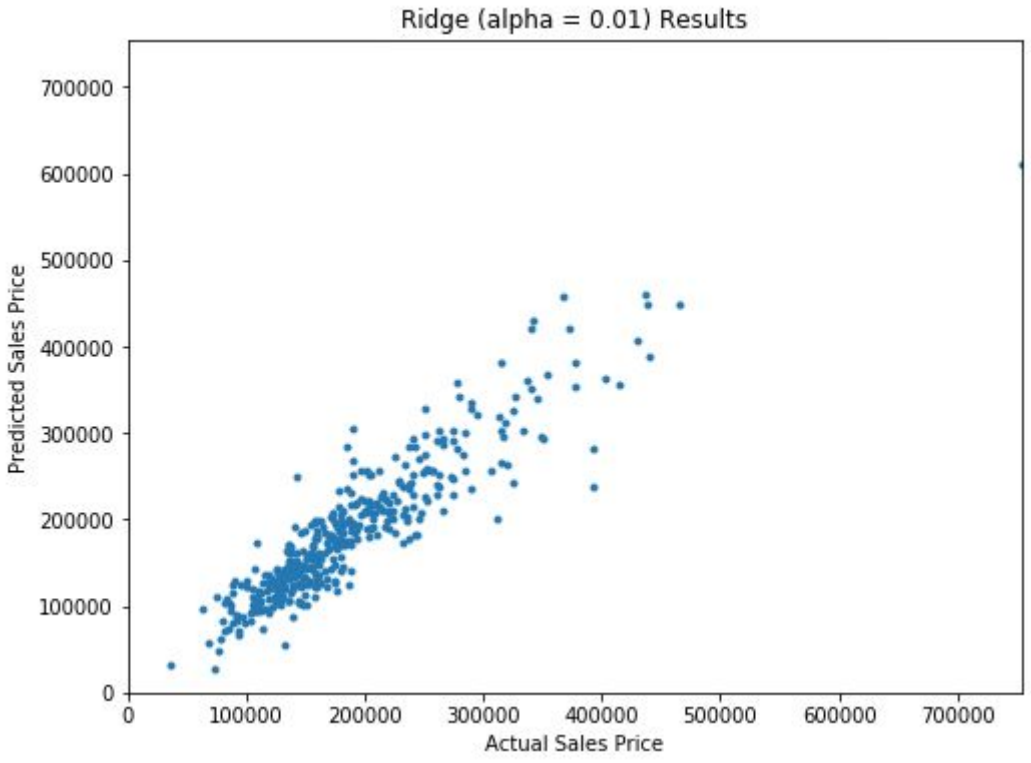


Ridge (alpha = 0.01)

35	YrSold	-973.891972
222	PoolQC_Gd	-203.107150
170	KitchenQual_Gd	-130.812094
16	HalfBath	-123.168535
17	KitchenAbvGr	-112.560256
...	...	...
0	1stFlrSF	315.903990
27	OverallQual	343.637940
34	YearRemodAdd	344.800558
33	YearBuilt	351.253869
15	GrLivArea	364.769793

From the above coefficients, we observe that the range/magnitude of the values is much lower than LASSO (alpha = 0.1).

Sum of coefficients: 1653692.182882933
Total number of coefficients: 31877
Number of coefficients not equal to 0: 19638
Ridge R2 for alpha = 0.01 is: 0.744802099960274



Ridge naturally gives us a much better R2 and therefore fit, even with a low value of 0.01 for alpha.

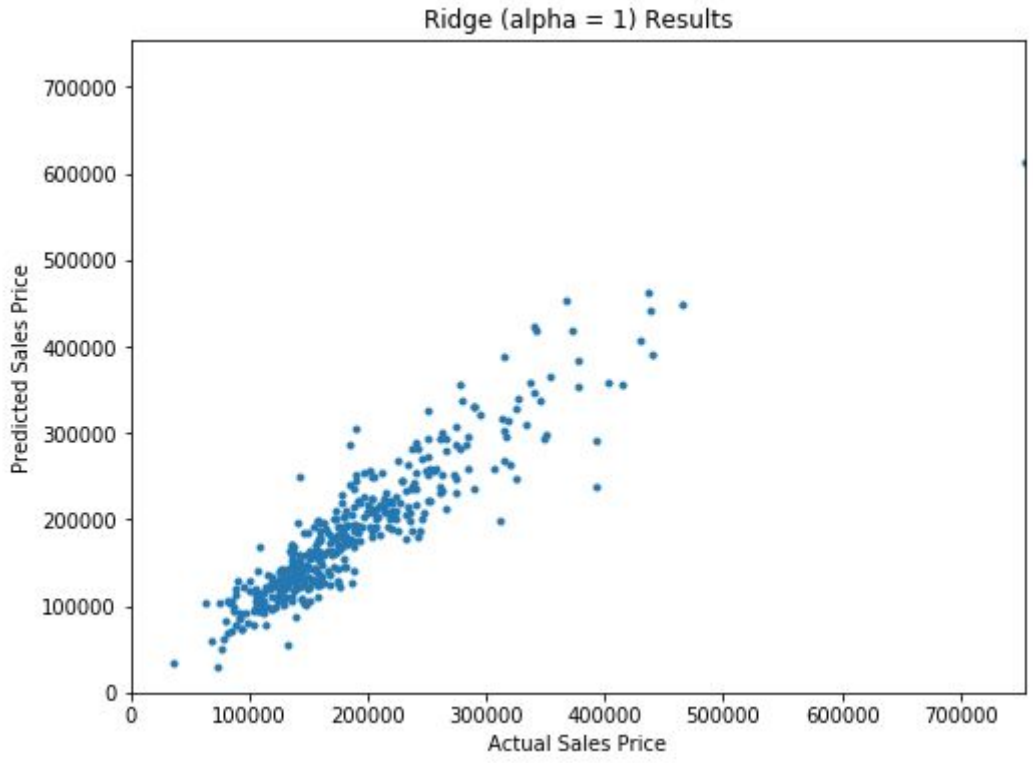
Ridge (alpha = 1)

35	YrSold	-608.356685
222	PoolQC_Gd	-203.441379
170	KitchenQual_Gd	-129.765402
17	KitchenAbvGr	-113.108723
61	BsmtQual_Gd	-106.933447
...	...	...
27	OverallQual	295.577809
0	1stFlrSF	304.433071
33	YearBuilt	350.711995
15	GrLivArea	363.586329
34	YearRemodAdd	377.039560

In comparison to Ridge (alpha = 0.01), as expected, we observe that the coefficients have a lower range/magnitude.

Sum of coefficients: 1576204.7309504985
Total number of coefficients: 31877
Number of coefficients not equal to 0: 19638
Ridge R2 for alpha = 1 is: 0.7588244202226753

We observe a better fit due to the higher R2 score but the number of non zero coefficients have remained the same.

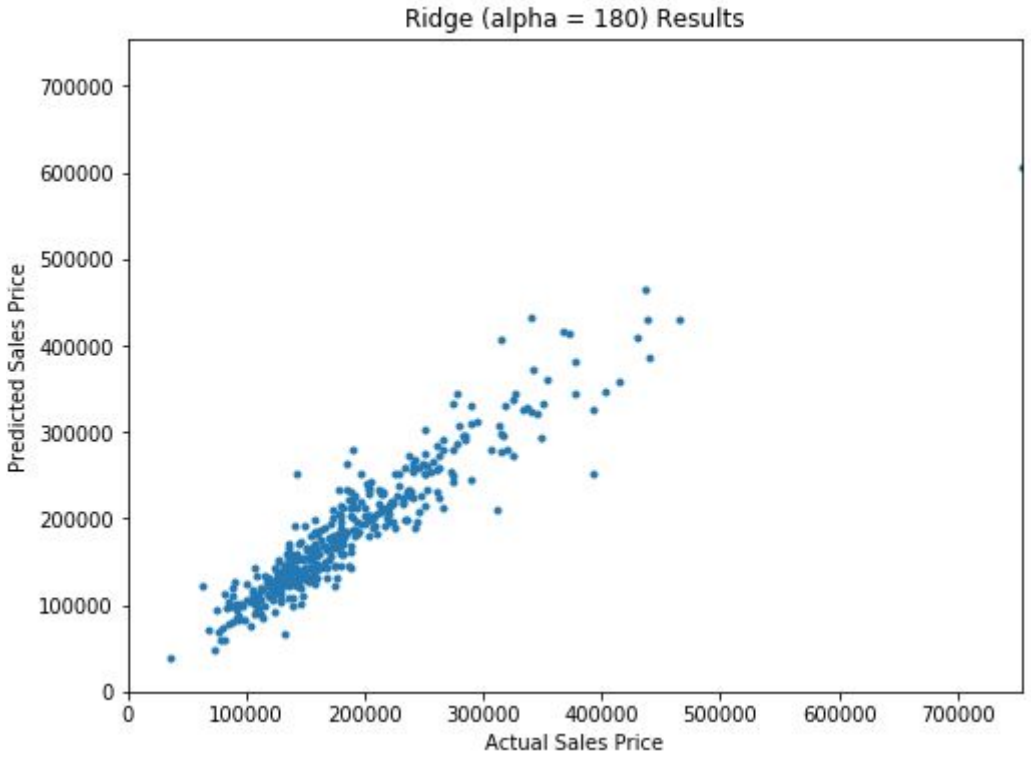


<u>Cross Val Score - Parameter Tuning : Ridge (alpha = 180)</u>

Performing cross validation from importing RidgeCV a few times has given us an alpha of 240. The summary obtain is below:

Sum of coefficients: 1049905.052453129
Total number of coefficients: 31877
Number of coefficients not equal to 0: 19638
Ridge R2 for alpha = 180 is: 0.8153834664576644

We observe that the R2 score has improved. This can be observed visually from the below scatter plot as most of the points are concentrated on a much straighter line with a reduced spread.



Conclusion

From our analysis we conclude that the most appropriate model for our dataset is the Ridge regression with a higher penalty parameter alpha. This is based on the better R2 score in comparison to the LASSO. However, if feature selection is of priority, then LASSO with higher penalty parameter alpha should be used.

Further Steps:

As an extension to the data analysis, we may have the below as new objectives:

- Perform Grid Search CV to achieve further parameter tuning.
- Produce models using alternative regression techniques such as:
    - Support Vector Machine (Linear, Polynomial, Gaussian RBF kernels)
    - Random Forest Regression
    - XG Boost
- Gather more data as time goes on for further model performance analysis.
- Using Principal Component Analysis for feature selection and explain the variance ratio i.e. relevance of each feature for a simpler model.