# ADVANCED LINEAR REGRESSION

## Assignment Part-II

Submitted by: SOUMAYADEEP MANNA

GitHub Repository link of the Assignment:
https://github.com/Soumayad96/House_Price_Prediction

SUBJECTIVE QUESTIONS AND ANSWERS:

### Question 1

***What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?***

Answer:

Ridge regression optimal alpha values is: **4.0**
Lasso regression optimal alpha values is: **0.0008**

As asked, we have doubled the alpha values now,

Ridge alpha is: **8.0** and Lasso alpha is: **0.0016**

After fitting the model with new alpha and predicting on testing data following changes is observed:

1. *Coefficients values changes for both the models.*
2. *In case of top features, first 3-4 features remained same after that few new features added or removed or order of importance changes. We don't see any random changes in top features list.*

**Ridge Model:**

**Top Features Before:**

1. GrLivArea -- (Ground living area ) -- higher the area higher the price
2. OverallQual -- (Overall house quality) -- better the quality higher the price
3. OverallCond -- (House Condition) -- better the house condition higher the price
4. Condition2_PosN -- (Nearby offsite area like park etc. ) -- houses having less off site area have lower price
5. RoofMatl_WdShngl -- (Roof material is Wood Shingles) -- houses with wood shingles have higher prices

6. Neighborhood_StoneBr -- (Stone brook neighborhood) -- Stone brook neighborhood have high house price
7. 1stFlrSF -- (1st Floor area) -- more the area higher the price
8. Neighborhood_Edwards -- (Edwards Neighborhood ) -- Edwards Neighborhood has lower house price
9. Neighborhood_Crawfor -- (Crawford Neighborhood ) -- Crawford Neighborhood has higher house price
10. Neighborhood_NoRidge -- (Northridge Neighborhood ) -- Northridge Neighborhood has lower house price

## Top Features After:

1. GrLivArea -- (Ground living area )
2. OverallQual -- (Overall house quality)
3. OverallCond -- (House Condition)
4. 1stFlrSF-- (1st floor area in square feet)
5. Neighborhood_Crawfor -- (Crawford neighborhood)
6. Neighborhood_Edwards -- (Edwards neighborhood )
7. Neighborhood_StoneBr -- (Stone brook neighborhood)
8. Neighborhood_NridgHt -- (Northridge Heights Neighborhood )
9. Neighborhood_NoRidge -- (Northridge Neighborhood )
10. GarageCars -- (Garage capacity car quantity wise)

## Lasso Model:

## Top Features Before:

1. GrLivArea -- (Ground living area ) -- higher the area higher the price
2. OverallQual -- (Overall house quality) -- better the quality higher the price
3. OverallCond -- (House Condition) -- better the house condition higher the price
4. GarageCars -- ( Garage size based on car capacity ) -- higher garage capacity have higher house prices
5. Neighborhood_StoneBr -- (Stone brook neighborhood ) -- Stone brook neighborhood have high house price
6. Neighborhood_NridgHt -- (Northridge Heights neighborhood ) -- Northridge Heights neighborhood have high house price
7. Neighborhood_NoRidge -- (Northridge neighborhood) -- Northridge neighborhood have high house price
8. Neighborhood_Somerst -- (Somerset Neighborhood ) -- Somerset Neighborhood have high house price
9. Neighborhood_Edwards -- (Edwards Neighborhood ) -- Edwards Neighborhood has lower house price
10. SaleCondition_Partial -- (New under built house ) -- New under built house has higher house price

## Top Features After:

1. GrLivArea -- (Ground living area )
2. OverallQual -- (Overall house quality)
3. OverallCond -- (House Condition)
4. GarageCars -- ( Garage size based on car capacity )
5. Neighborhood_NridgHt -- (Northridge Heights neighborhood )
6. TotalBsmtSF -- (Total basement area in square foot)
7. Neighborhood_Edwards -- (Edwards Neighborhood )
8. Neighborhood_Somerst -- (Somerst Neighborhood )
9. Neighborhood_Crawfor -- (Crawford Neighborhood )
10. SaleCondition_Partial -- (New under built house

Though there is change in coefficient values but there is not random change in importance of the features.

**Question 2**

***You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?***

Answer:

We have used regularization techniques to implement the Ridge and Lasso regression models on our data. After applying cross validation and model fitting, we calculated model metrics and importance of the coefficients. Following is observed:

1. Lasso model performed better than Ridge model on train and test data on a slight margin. Though we cannot say that Ridge model is not performed well.
2. Lasso model have better R2 score on test data, RSS and MSE error term are less on test data compared to Ridge Model.
3. Lasso model optimizes the features and trimmed out 88 top important features from total of 218 features, which increases the model efficiency.
4. Optimal alpha for ridge is 4.0 and for lasso is 0.0008. While doubling the alpha values we seen that model performance is decreasing which explains higher lambda value overfits the model.

**So, let's consider the Lasso Model from our analysis, where the following details are as follows:**

1. Optimal alpha value is 0.008, which is also known as hyper parameter in case of regularization techniques.
2. It explains 90% of variance present in the data correctly when tested on unknown data.

**Top features which helps to predict or control the Target variable (Sale Price) is as follows:**

1. GrLivArea -- (Ground living area ) -- higher the area higher the price
2. OverallQual -- (Overall house quality) -- better the quality higher the price
3. OverallCond -- (House Condition) -- better the house condition higher the price
4. GarageCars -- ( Garage size based on car capacity ) -- higher garage capacity have higher house prices
5. Neighborhood_StoneBr -- (Stone brook neighborhood ) -- Stone brook neighborhood have high house price
6. Neighborhood_NridgHt -- (Northridge Heights neighborhood ) -- Northridge Heights neighborhood have high house price
7. Neighborhood_NoRidge -- (Northridge neighborhood) -- Northridge neighborhood have high house price
8. Neighborhood_Somerst -- (Somerset Neighborhood ) -- Somerset Neighborhood have high house price
9. Neighborhood_Edwards -- (Edwards Neighborhood ) -- Edwards Neighborhood has lower house price
10. SaleCondition_Partial -- (New under built house ) -- New under built house has higher house price|

**Question 3**

***After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?***

Answer:

**Based on first Lasso model where alpha = 0.0008 we found that top 5 features are as follows:**

*1. GrLivArea -- (Ground living area ) -- higher the area higher the price*
*2. OverallQual -- (Overall house quality) -- better the quality higher the price*
*3. OverallCond -- (House Condition) -- better the house condition higher the price*
*4. GarageCars -- ( Garage size based on car capacity ) -- higher garage capacity have higher house prices*
*5. Neighborhood_StoneBr -- (Stone brook neighborhood ) -- Stone brook neighborhood have high house price*

**So we have removed following 5 features from train and test data, and re build the model using same alpha values.**

**Now the top 5 features are as follows:**
*1. BsmtUnfSF -- (Unfinished basement area)*
*2. HalfBath -- (Half baths above grade)*
*3. Neighborhood_BrDale -- (Briardale Neighborhood)*
*4. 2ndFlrSF -- (2nd Floor Area in Sqrft )*
*5. Neighborhood_NridgHt -- (Northridge Heights neighborhood )*

**This explains that when we removed the first top 5 features and build the model again, correlation between features are changed and we get a totally new set of top 5 important features for the model.**

**Question 4**

***How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why***?

Answer:

A model needs to be made robust and generalizable so that they are not impacted by outliers in the training data. The model should also be generalisable so that the test accuracy is not lesser than the training score. The model should be accurate for datasets other than the ones which were used during training. Too much weightage should not be given to the outliers so that the accuracy predicted by the model is high.

To ensure that this is not the case, the outlier analysis needs to be done and only those which are relevant to the dataset need to be retained. Those outliers which it does not make sense to keep must be

removed from the dataset. This would help increase the accuracy of the predictions made by the model. Confidence intervals can be used (typically 3-5 standard deviations). This would help standardize the predictions made by the model. If the model is not robust, it cannot be trusted for predictive analysis.

==========