SURPRISE HOUSING ASSIGNMENT

SUBJECTIVE QUESTIONS

SHRUTHIP VENKATESH
ML-C54

Subjective Questions – Assignment Part II

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:

Check here for detailed analysis under section *Subjective → Question 1*

Ridge:

Optimal value for Ridge model is **8**.

On doubling the value of alpha, which is **16**, the following observations are made.

1) In train data, there is a negligible drop of 0.01 in R2 score and a negligible increase of 0.001 in RMSE values.

Metric	Ridge Regression	Ridge Double Regression
R2 Score (Train)	**0.906306**	**0.896690**
R2 Score (Test)	0.892148	0.888828
RMSE (Train)	**0.033440**	**0.035114**
RMSE (Test)	0.034280	0.034803

2) There is a negligible change in the coefficients of the features. Below table shows an example.

Feature	Ridge	Ridge Double
LotArea	0.022978	0.016849
OverallQual	0.052516	0.044336
OverallCond	0.030442	0.022176
YearBuilt	0.003986	0.002515
YearRemodAdd	0.007143	0.007820

3) There is no change in the number of predictors

4) The top **6** predictors in order [*'OverallQual', 'GrLivArea', '1stFlrSF', 'TotRmsAbvGrd', 'Neighborhood_NoRidge', 'Total_floor_SF'*] remain the same even after doubling the alpha value.

5) Irrespective of the order of the predictors, **18** predictors out of top **20** predictors remain the same even after doubling the alpha value.

	Ridge Features	Ridge Double Features
0	OverallQual	OverallQual
1	GrLivArea	GrLivArea
2	1stFlrSF	1stFlrSF
3	TotRmsAbvGrd	TotRmsAbvGrd
4	Neighborhood_NoRidge	Neighborhood_NoRidge
5	Total_floor_SF	Total_floor_SF
6	Neighborhood_StoneBr	TotalBsmtSF
7	2ndFlrSF	KitchenQual_Ex
8	TotalBsmtSF	2ndFlrSF
9	KitchenQual_Ex	Neighborhood_StoneBr
10	OverallCond	Fireplaces
11	GarageCars	GarageCars
12	ExterQual_Ex	ExterQual_Ex
13	BsmtFinSF1	BsmtQual_Ex
14	Fireplaces	BsmtFinSF1
15	BsmtQual_Ex	BsmtExposure_Gd
16	Neighborhood_NridgHt	GarageArea
17	MasVnrArea	OverallCond
18	BsmtUnfSF	Neighborhood_NridgHt
19	LotArea	BsmtUnfSF

Lasso:

Optimal value for Lasso model is **0.0001**.

On doubling the value of alpha, which is **0.0002**, the following observations are made.

1) In train data, there is a negligible drop of 0.01 in R2 score and a negligible increase of 0.001 in RMSE values.

Metric	Lasso Regression	Lasso Double Regression
R2 Score (Train)	**0.909805**	**0.901033**
R2 Score (Test)	0.901734	0.901742
RMSE (Train)	**0.032810**	**0.034368**
RMSE (Test)	0.032721	0.032719

2) There is a negligible change in the coefficients of the features. Below table shows an example.

Feature	Lasso	Lasso Double
LotArea	0.037430	0.018370
OverallQual	0.078478	0.089601
OverallCond	0.045464	0.037895
YearBuilt	0.000000	0.000000
YearRemodAdd	0.001666	0.001057

3) Number of predictors dropped to **93** from *132*.

4) The top **4** predictors in order [*'GrLivArea', 'OverallQual', 'TotalBsmtSF', 'Neighborhood_StoneBr'*] remain the same even after doubling the alpha value.

5) Irrespective of the order of the predictors, **18** predictors out of top **20** predictors remain the same even after doubling the alpha value.

	Lasso Features	Lasso Double Features
0	GrLivArea	GrLivArea
1	OverallQual	OverallQual
2	TotalBsmtSF	TotalBsmtSF
3	Neighborhood_StoneBr	Neighborhood_StoneBr
4	GarageQual_Ex	Neighborhood_NoRidge
5	Neighborhood_NoRidge	KitchenQual_Ex
6	OverallCond	OverallCond
7	TotRmsAbvGrd	TotRmsAbvGrd
8	KitchenQual_Ex	ExterQual_Ex
9	ExterQual_Ex	GarageCars
10	LotArea	Neighborhood_NridgHt
11	Neighborhood_NridgHt	BsmtQual_Ex
12	GarageCars	BsmtExposure_Gd
13	BsmtQual_Ex	BsmtFinSF1
14	Exterior1st_BrkFace	Exterior1st_BrkFace
15	BsmtExposure_Gd	Neighborhood_Crawfor
16	Neighborhood_Crawfor	Fireplaces
17	BsmtFinSF1	BldgType_1Fam
18	ScreenPorch	ScreenPorch
19	Functional_Typ	LotArea

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:

Let's have a look at the metrics of both ridge and lasso regression models,

Metric	Ridge Regression Alpha = 8	Lasso Regression Alpha = 0.0001
R2 Score (Train)	0.906306	0.909805
R2 Score (Test)	0.892148	0.901734
RMSE (Train)	0.033440	0.032810
RMSE (Test)	0.034280	0.032721

- Since it is evident that both the models *R2 score* is same around **0.90** for both train and test data and *RMSE value* is same around **0.033** for both train and test data, it is better to select a model which is **simple**.
- In that terms, **lasso model** does a better job since it does feature selection which resulted in *132 features* whereas ridge model has *278 features* which is 146 features more than lasso model.

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:

Check here for detailed analysis under section *Subjective → Question 3*

- The top 5 predictors of the Lasso model is [*'GrLivArea', 'OverallQual', 'TotalBsmtSF', 'Neighborhood_StoneBr', 'GarageQual_Ex'*]

Upon dropping the above mentioned features, here are the observations,

1) The optimal alpha value remained same to be 0.0001

2) In train data, there is a negligible drop of 0.01 in R2 score and a negligible increase of 0.001 in RMSE values.

Metric	Lasso Regression	Lasso Regression after drop
R2 Score (Train)	**0.909805**	**0.902314**
R2 Score (Test)	0.901734	0.885963
RMSE (Train)	**0.032810**	**0.0341451**
RMSE (Test)	0.032721	0.0341451

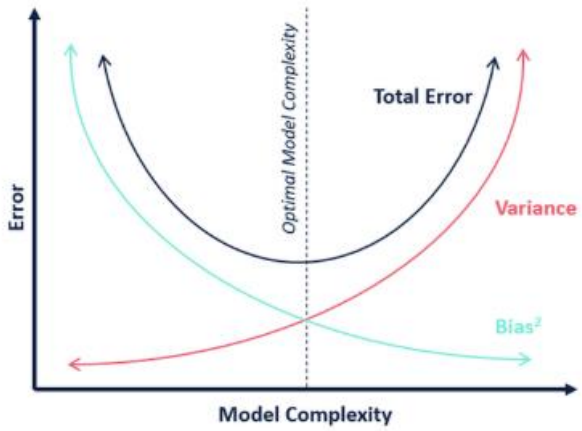
3) The new top 5 predictors in order are,
    1) Total_floor_SF
    2) 1stFlrSF
    3) 2ndFlrSF
    4) OverallCond
    5) BsmtUnfSF

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

Answer:

Let's understand few terms,

- *Bias* : This refers on how accurate the model likely to be on future (test) data
- *Variance* : This refers to the degree of changes in the model itself with respect to changes in the training data.



- For a model to be **robust**, it should not show significant change in performance for change in data, which means the variance should be low. As seen above in the graph, a **simple** model will have low variance but the trade off is bias will be high. If the model becomes so simple, the accuracy will be low.
- For a model to be **general**, it should not overfit the data. It should perform equally good in train and test data, which means the bias should be low. As seen above in the graph, a **complex** model will have low bias but the trade off is variance will be high. If the model becomes so complex, the model will memorize the data.
- From the above two statements, we could see that we need to have a balance between bias and variance for achieving a robust and generalized model. This is called bias-variance tradeoff.

- In order to build a robust and generalizable model, we have the concept of regularization to optimally simplify models.
- For regression techniques, it is achieved by adding a regularization term to the cost function that adds up the absolute values (Lasso) or the squares (Ridge) of the parameters of the model.

Implications on accuracy:

- Regularization, significantly reduces the variance of the model, without substantial increase in its bias.
- When we add regularization, we're modifying the loss function to penalize large coefficients, which distracts from the goal of optimizing accuracy. The larger the regularization penalty, the more we deviate from our goal of optimizing training accuracy. Hence, training accuracy decreases.
- Even though the training accuracy goes down, since the model is becoming generic, the test accuracy increases.
- Upon building an optimal model, we arrive at an accuracy which is neither too high nor too low.