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 **3.3333333333333335**.

On doubling the value of alpha, which is **6.666666666666667**, the following observations are made.

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

Metric	Ridge Regression	Ridge Double Regression
R2 Score (Train)	0.938550	0.933284
R2 Score (Test)	0.897946	0.897809
RMSE (Train)	0.032111	0.033458
RMSE (Test)	0.040918	0.040945

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

Feature	Ridge	Ridge Double
LotArea	0.037856	0.029977
OverallQual	0.092293	0.077689
OverallCond	0.052950	0.043672
YearBuilt	0.016924	0.011705
YearRemodAdd	0.013163	0.013851

3) There is no change in the number of predictors

4) The top **3** predictors in same order [*'OverallQual', '1stFlrSF', 'GrLivArea'*] remain the same and the following **7** predictors in no order *['OverallCond', 'Total_floor_SF', 'GarageArea', 'TotRmsAbvGrd', 'TotalBsmtSF', 'Neighborhood', '2ndFlrSF']* remain the same even after doubling the alpha value.

5) Irrespective of the order of the predictors, **19** predictors out of top **20** predictors remain the same even after doubling the alpha value.

	Ridge Features	Ridge Double Features
0	OverallQual	OverallQual
1	1stFlrSF	1stFlrSF
2	GrLivArea	GrLivArea
3	OverallCond	Total_floor_SF
4	Total_floor_SF	GarageArea
5	GarageArea	OverallCond
6	TotRmsAbvGrd	TotRmsAbvGrd
7	2ndFlrSF	Neighborhood_Crawfor
8	Neighborhood_Crawfor	2ndFlrSF
9	TotalBsmtSF	TotalBsmtSF
10	LotArea	GarageCars
11	Neighborhood_StoneBr	Neighborhood_StoneBr
12	GarageCars	LotArea
13	BedroomAbvGr	Total_Bathrooms
14	Neighborhood_NridgHt	BedroomAbvGr
15	SaleType_ConLD	Neighborhood_NridgHt
16	Fireplaces	Exterior1st_BrkFace
17	Exterior1st_BrkFace	HalfBath
18	Total_Bathrooms	Fireplaces
19	HalfBath	Neighborhood_NoRidge

Lasso:

Optimal value for Lasso model is **0.00015777777777777776**.

On doubling the value of alpha, which is **0.00031555555555555555**, the following observations are made.

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

Metric	Lasso Regression	Lasso Double Regression
R2 Score (Train)	0.931790	0.921966
R2 Score (Test)	0.910902	0.911047
RMSE (Train)	0.033831	0.036185
RMSE (Test)	0.038232	0.038201

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

Feature	Lasso	Lasso Double
LotArea	0.035436	0.016681
OverallQual	0.147558	0.159259
OverallCond	0.066028	0.054041
YearBuilt	0.000000	0.000000
YearRemodAdd	0.000000	0.000000

3) Number of predictors dropped to **88** from *123*.

4) The top **2** predictors in order [*‘GrLivArea’, ‘OverallQual’*] remain the same and the following **6** predictors in no order *['TotalBsmtSF','Total_floor_SF','GarageArea','OverallCond','Total_Bathrooms','Neighborhood’]* remain the same.

5) Irrespective of the order of the predictors, **19** 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	Total_floor_SF	TotalBsmtSF
3	OverallCond	Total_floor_SF
4	TotalBsmtSF	GarageArea
5	GarageArea	OverallCond
6	Neighborhood_Crawfor	Total_Bathrooms
7	Total_Bathrooms	Neighborhood_Crawfor
8	LotArea	GarageCars
9	GarageCars	Total_porch_sf
10	Neighborhood_StoneBr	TotRmsAbvGrd
11	Fireplaces	Fireplaces
12	Neighborhood_NridgHt	Neighborhood_NridgHt
13	Total_porch_sf	BsmtQual_Ex
14	TotRmsAbvGrd	Exterior1st_BrkFace
15	SaleType_ConLD	LotArea
16	Exterior1st_BrkFace	Neighborhood_StoneBr
17	Neighborhood_NoRidge	Neighborhood_NoRidge
18	BsmtQual_Ex	BsmtExposure_Gd
19	Functional_Typ	Functional_Typ

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 = 3.33	Lasso Regression Alpha = 0.00015
R2 Score (Train)	**0.938550**	**0.931790**
R2 Score (Test)	0.897946	0.910902
RMSE (Train)	**0.032111**	**0.033831**
RMSE (Test)	0.040918	0.038232

- Since it is evident that both the models *R2 score* is same around **0.93** for train data and same around **0.90** for test data and *RMSE value* is same around **0.033** for train data and same around **0.038** for test data, it is better to select a model which is **simple**.
- In that terms, **lasso model** does better job since it does feature selection which resulted in *123 features* whereas ridge model has *298 features* which is 175 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', Total_floor_SF ', OverallCond ', TotalBsmtSF'*]

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 in R2 score and a negligible increase of 0.001 in RMSE values.

Metric	Lasso Regression	Lasso Regression after drop
R2 Score (Train)	**0.931790**	**0.930008**
R2 Score (Test)	0.910902	0.890314
RMSE (Train)	**0.033831**	**0.034269**
RMSE (Test)	0.038232	0.042420

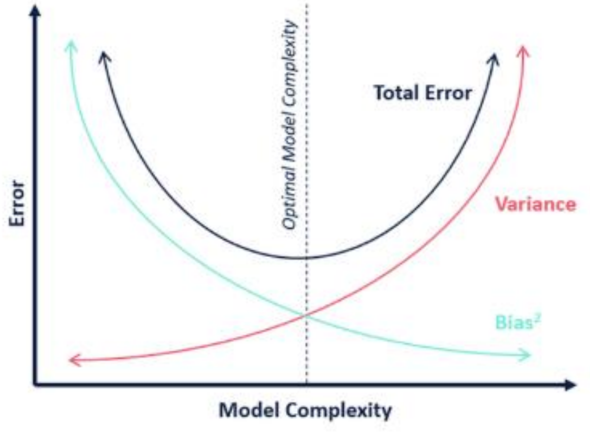
3) The new top 5 predictors in order are,
   1) 1stFlrfloorSF
   2) 2ndFlrSF
   3) GarageArea
   4) Neighborhood
   5) SaleType

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.