SUBJECTIVE QUESTIONS ASSGINMENT PART – II

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: -

The Optimal value of alpha for Ridge regression is Alpha = 2

The Optimal value of alpha for Lasso regression is Alpha = 0.01

Now let's see the model for present Alpha value and doubled Alpha value

Ridge regression: -

For Alpha = 2

R2_Score(train) = 0.9482290910432771
R2_Score(test) = 0.8884603744201733
RMSE = 0.12896746935928943

	Variable	Coef
0	constant	12.027
12	GrLivArea	0.124
3	OverallQual	0.107
8	TotalBsmtSF	0.049
20	GarageArea	0.045
4	OverallCond	0.041
19	Fireplaces	0.033
6	BsmtFinSF1	0.027
2	LotArea	0.016
13	BsmtFullBath	0.011
21	WoodDeckSF	0.004
15	FullBath	0.003
16	HalfBath	0.003

For Alpha = 4 (double the optimal)

R2_Score(train) = 0.9445861193262779
R2_Score(test) = 0.8890969699007841
RMSE = 0.1285989116082882

	Variable	Coef
0	constant	12.027
12	GrLivArea	0.124
3	OverallQual	0.107
8	TotalBsmtSF	0.049
20	GarageArea	0.045
4	OverallCond	0.041
19	Fireplaces	0.033
6	BsmtFinSF1	0.027
2	LotArea	0.016
13	BsmtFullBath	0.011
21	WoodDeckSF	0.004
15	FullBath	0.003
16	HalfBath	0.003

Here we can observe not much difference in the variables and coefficients but there is a slight decrease in R2_Score of train and test and a slight decrease in RMSE also.

Lasso Regression: -

For Alpha = 0.01

R2_Score(train) = 0.8910076106999829
R2_Score(test) = 0.8635391905839304
RMSE : 0.14264927159863097

For Alpha = 0.02

R2_Score(train) = 0.8793617901398336
R2_Score(test) = 0.855162028257104
RMSE : 0.1469625923622438

	Variable	Coeff		Variable	Coeff
0	constant	12.027	0	constant	12.027
1	MSSubClass	-0.017	1	MSSubClass	-0.017
2	LotArea	0.016	2	LotArea	0.016
3	OverallQual	0.107	3	OverallQual	0.107
4	OverallCond	0.041	4	OverallCond	0.041
6	BsmtFinSF1	0.027	6	BsmtFinSF1	0.027
8	TotalBsmtSF	0.049	8	TotalBsmtSF	0.049
12	GrLivArea	0.124	12	GrLivArea	0.124
13	BsmtFullBath	0.011	13	BsmtFullBath	0.011
15	FullBath	0.003	15	FullBath	0.003
16	HalfBath	0.003	16	HalfBath	0.003
18	KitchenAbvGr	-0.003	18	KitchenAbvGr	-0.003
19	Fireplaces	0.033	19	Fireplaces	0.033
20	GarageArea	0.045	20	GarageArea	0.045
21	WoodDeckSF	0.004	21	WoodDeckSF	0.004
28	property_age	-0.083	28	property_age	-0.083
30	remodel_age	-0.031	30	remodel_age	-0.031

Here, If we further increase the Alpha value then the feature selection will increase and Penalty in lasso forces some of the coefficients estimates to be exactly equal to zero.
Lasso performs better in situations where only a few among all the predictors that are used to build our model have significant influence on the response variable.
Ridge Performs better when all the variables have almost the same influence on the response v ariable.
So, we can conclude that higher the lambda more the regularization.

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: - Let's start comparing with RMSE's for both Ridge and Lasso.

RMSE for Ridge = 0.12896746935928943

RMSE for Lasso = 0.14264927159863097

Here, we can see that RMSE values are almost same with slight difference, where Ridge is Performing better in this case compared to lasso.

But Lasso help us with feature selection by reducing some of the features as Penalty in lasso forces some of the coefficients estimates to be exactly equal to zero.

By this we can easily predict the features using lasso.

So, I choose Lasso for prediction of my final model.

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: -

In the present model the 5 most important predictors are:

1) GrLivArea

2) OverallQual

3) TotalBsmtSF

4) property_age

5) remodel_age

After excluding this when I recreate the model the observations are:

The older values of R2 and RMSE are

0.8910076106999829
0.8635391905839304
RMSE : 0.14264927159863097

The New values of R2 and RMSE are :

0.8522390741586945
0.8167495560111402
RMSE : 0.16530580493693625

Here, we can observe the increase in the RMSE value and decrease in the R2 scores of both Train and test data. which is not acceptable. This helps us to understand the need of the previous columns in our model.
The new top 5 features are:
1) FullBath
2) OverallCond
3) MSZoning_RL
4) 2ndFlrSF
5) EnclosedPorch

Question 4

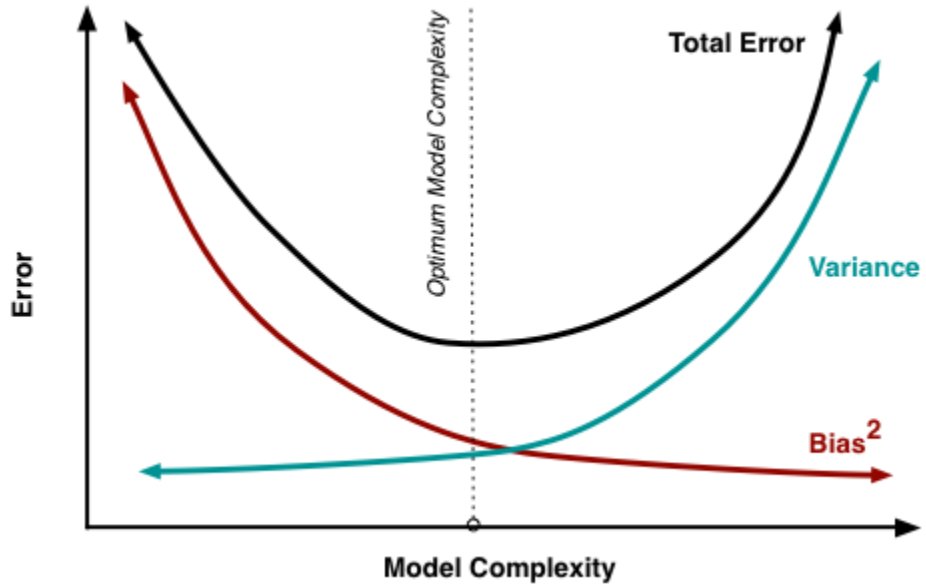
How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Answer: -

According to occam's razor a model should be as simple as possible but robust.

- The model should be generalized so the test accuracy isn't lesser than the training score.
- The model should be accurate for datasets apart from those which were used during training.
- An excessive amount of importance shouldn't be given to the outliers in order that the accuracy predicte d by the model is high.
- To make sure that this is often not the case, the outlier's analysis must be done and only those which are relevant to the dataset must be retained.
- Hence, Simpler models are more generic and more widely applicable.
- Here, simpler models requires less training samples compared to the complex models and these simple models are easier to train and effective also.
- While building complex models these models change widely when we change the training dataset
- Simple models are more robust compared to the complex models.
- Simple models have high bias and low variance.
- Complex models have high variance and low bias.
- Generally complex models lead to overfitting where they work very well for training samples but work very badly for test samples.
- Simple models make more errors in training set but fit nicely in test set much better than complex mode ls.
- But we should not make our models too much simple that it becomes unusable.
- Hence to make the models more robust and generalizable we must use simple models.
- Regularization is a process used to make the models simpler rather than making it complex.

- Here, bias specifies how much error the model likely to make in the test data.
- Here variance specifies how sensitive is the model to input data.
- Regularization is the process of deliberately simplifying models to achieve the correct balance between keeping the model simple and not too naïve.
- This improve accuracy and simplifies the model
- This provides us bias variance trade off.



Accuracy of the model can be maintained by keeping the balance between bias and variance as it minim ises the total error as shown in the graph above

- If we make complex models then we need to change the model for each and every small changes also.
- So, complex models becomes very unstable and extremely sensitive to any changes in the training data.
- This could not be the case in simple models. So, we prefer simple models which are robust too.

_____x_____x_____