Advanced Linear Regression assignment questions

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:
Optimal value of alpha for Ridge = 1
Optimal value of alpha for Lasso = 100

By doubling the value of alpha:

For Ridge : By keeping alpha = 2 for Ridge regression there is no significant change in the model behaviour. R-square is almost the same for train and test data. Root mean square error for train increased by 500 but it reduced by 100 for test data.

Metric	alpha = 1	alpha = 2
R-square-Train	0.946	0.943
RMSE-Train	14186.373	14576.307
R-square-Test	0.900	0.901
RMSE-Test	18938.397	18862.00

Top 5 predictor variables with positive coefficients:

Precedence	alpha = 1	alpha = 2
1	GrLivArea	GrLivArea
2	TotalBsmtSF	TotalBsmtSF
3	1stFlrSF	1stFlrSF
4	YearBuilt	OverallQual_9 (Excellent Quality)
5	OverallQual_9 (Excellent Quality)	YearBuilt

Top 5 predictor variables with negative coefficients:

Precedence	alpha = 1	alpha = 2
1	BsmtUnfSF (Unfinished square feet of basement area)	BsmtUnfSF
2	Functional_Sev (Severely damaged home functionality)	MSZoning_C
3	MSZoning_C (all) (Commercial zone)	OverallQual_3
4	OverallQual_3 (material and finish of the house - Fair)	OverallCond_3
5	OverallCond_3 (condition of the house - Fair)	Functional_Sev

For Lasso : By keeping alpha = 200 for Lasso regression there is a slight change in the model behaviour. R-square value has declined a little bit compared to the values when alpha = 100. Root mean square error also has been increased for both train and test dataset.

Metric	alpha = 100	alpha = 200
R-square-Train	0.931	0.917
RMSE-Train	16000.584	17486.85
R-square-Test	0.908	0.901
RMSE-Test	18196.218	18866.73

Top 5 predictor variables with positive coefficients:

Precedence	alpha = 100	alpha = 200
1	GrLivArea	GrLivArea
2	TotalBsmtSF	TotalBsmtSF
3	OverallQual_9 (Excellent quality)	OverallQual_9
4	YearBuilt	OverallQual_8

5	OverallQual8 (Very good quality)	Neighborhood_Crawfor (Crawford neighborhood)

Top 5 predictor variables with negative coefficients:

Precedence	alpha = 100	alpha = 200
1	BsmtUnfSF (Unfinished square feet of basement area)	OverallCond_3
2	OverallCond_3 (condition of the house - Fair)	ExterQual_TA (Average quality of the material on exterior)
3	MSZoning_C (Commercial zone)	SaleCondition_Abnorml (Abnormal sale condition)
4	OverallQual_3 (material and finish of the house - Fair)	OverallCond_4
5	OverallCond_4 (Below Average)	GarageCars_0 (no option of garage cars)

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 :
From an accuracy point of view Ridge and Lasso both are giving almost similar results when comparing the train and test r-square values. In terms of errors also both the models are performing pretty decent and errors are almost equal.
But in this case Lasso would be preferred over Ridge and the key reason for that is the feature elimination done by Lasso. Here we fed the model with 310 features and out of these 310 features Lasso has made beta coefficient to zero for 201 features. So if we continue with Lasso we only need to use 109 features making the model better.

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:
The five most important predictor variables found with Lasso:

1. GrLivArea
2. TotalBsmtSF
3. OverallQual_9
4. YearBuilt
5. OverallQual8

After removing these features from the train and test data built the model again using Lasso with similar value of alpha (100)

The five most important predictor variables now are as follows:

1. BsmtUnfSF
2. 1stFlrSF
3. Neighborhood_MeadowV
4. Neighborhood_Blueste
5. Electrical_FuseP

Question 4

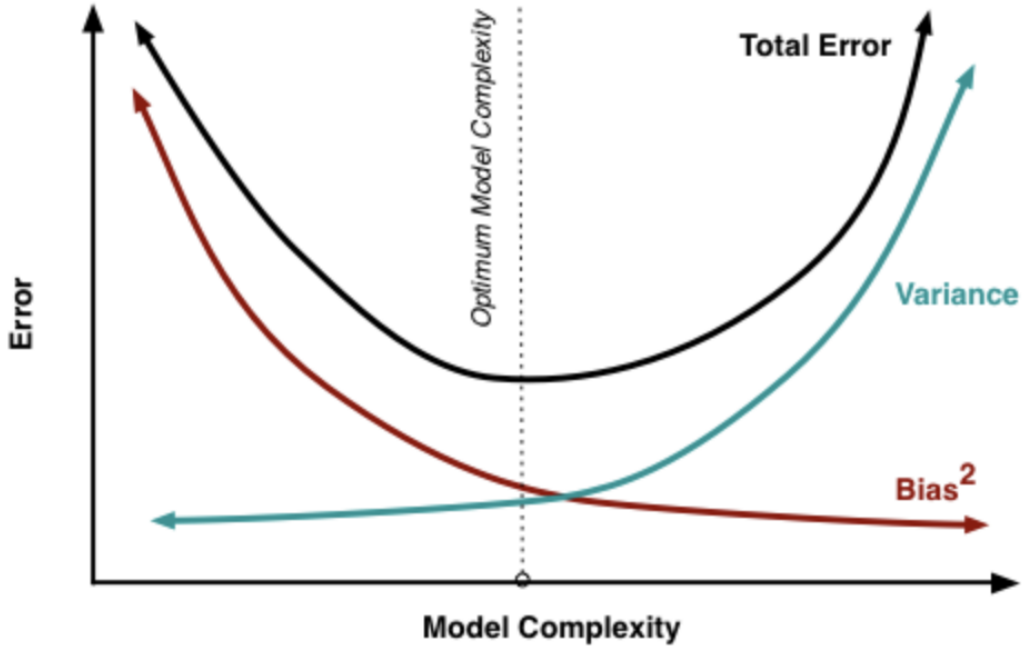
How can you make sure that a model is robust and generalisable? What are the

implications of the same for the accuracy of the model and why?

Answer:
In order to make the model robust and generalisable we need to make sure that the model is not too complex but also need to ensure that the model must not be very simple. When a model is complex it memorises the training points and gives very good accuracy on training data but fails to perform on unseen data , which signifies that the variance is high and the model is overfit. To reduce models from being overfit we use regularisation. Ridge and Lasso are the regularisation techniques used in Linear regression which penalises the model for using higher beta coefficients.

Using regularisation we are allowing some bias to be added into the model as a tradeoff to reduce variance.

Bias-var tradeoff:



As model's complexity decreases , variance in the model also decreases but bias increases. To find out the trade off between bias and variance we need to using hyperparameter tuning to find the best hyperparam where the model has lowest possible variance and bias.

In case of Ridge and Lasso Regression alpha(lambda) is the hyperparameter , when alpha is low which means regularisation is very low meaning very low penalty on the higher coefficients hence variance in the model is high but bias is low. As the value of alpha increases, the model becomes more generalised and robust with decrease in variance but it also adds some bias into the model as a trade off.