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?

ANS:

Point 1: Optimal values of alpha

The optimal value of alpha for

Ridge = 10

Lasso = 0.01

Point 2:

If we double alpha for **Ridge regression:**

alpha = 10

i.e. double alpha = 20
then

R2 of train 0.9229513950861605

R2 of test 0.914055331097043

RMSE is: 0.014458590875745816

Ridge parameters and its coeff:

(Only few are mentioned here)

[('constant', 11.93), ('OverallQual', 0.067), ('Neighborhood_StoneBr', 0.063), ('Neighborhood_Crawfor', 0.055), ('Exterior1st_BrkFace', 0.045), ('SaleCondition_Normal', 0.044), ('Neighborhood_NridgHt', 0.043), ('Condition2_Norm', 0.043), ('OverallCond', 0.04), ('GrLivArea', 0.04), ('Condition1_Norm', 0.039), ('MSZoning_FV', 0.036), ('SaleCondition_Partial', 0.03), ('1stFlrSF', 0.029), ('TotalBsmtSF', 0.027), ('Neighborhood_BrkSide', 0.027), ('GarageArea', 0.025), ('Exterior2nd_Wd Sdng', 0.025), ('TotRmsAbvGrd', 0.024), ('Neighborhood_ClearCr', 0.024), ('BsmtFinSF1', 0.022), ('LotConfig_CulDSac', 0.022), ('Exterior1st_MetalSd', 0.022), ('MSZoning_RL', 0.021), ('Neighborhood_Somerst', 0.021), ('SaleType_New', 0.021), ('2ndFlrSF', 0.02), ('Neighborhood_Veenker', 0.02), ('FullBath', 0.019), ('MonthSold_September', 0.019), ('Neighborhood_NoRidge', 0.018), ('Fireplaces', 0.017), ('MSSubClass_2-STORY 1945 & OLDER', 0.017), ('HalfBath', 0.016), ('GarageCars', 0.016), ('GarageType_Basment', 0.016), ('ScreenPorch', 0.015), ('MSSubClass_1-1/2 STORY FINISHED ALL AGES', 0.015), ('Exterior1st_VinylSd', 0.015), ('Exterior2nd_MetalSd', 0.015), ('Foundation_PConc', 0.015), ('MonthSold_July', 0.015), ('HeatingQC', 0.014), ('GarageQual', 0.014), ('MonthSold_May', 0.014), ('WoodDeckSF', 0.013), ('LotFrontage', 0.012), ('BsmtFinType1', 0.012), ('BsmtFullBath', 0.012), ('KitchenQual', 0.012), ('MasVnrType_Stone', 0.012), ('Functional', 0.011), ('Condition1_RRAn', 0.011), ('SaleType_ConLD', 0.011), ('LotArea', 0.01), ('BsmtQual', 0.01), ('MSSubClass_1-STORY 1946 & NEWER ALL STYLES', 0.01), ('MSSubClass_SPLIT FOYER', 0.01), ('Exterior2nd_HdBoard', 0.01), *Top 5 parameters:*

('OverallQual', 0.067)
('Neighborhood_StoneBr', 0.063)
('Neighborhood_Crawfor', 0.055)
('Exterior1st_BrkFace', 0.045)
('SaleCondition_Normal', 0.044)

Final Predicted house price mean

191061.2087084514

Observations:

- R2 on train and test decreases
- RMSE decreases
- Value of coeffiecients decreased
- Mean of the final predicted sale price decreased.

LASSO

alpha = 0.001

doubled alpha = 0.002

R2 of train 0.9032813073216734

R2 of test 0.9134095631370859

RMSE is: 0.014567229315486927

Lasso parameters and its coeff:

(Only few are mentioned here)

('constant', 12.023), ('OverallQual', 0.081), ('GrLivArea', 0.073), ('OverallCond', 0.04), ('TotalBsmtSF', 0.035), ('SaleCondition_Partial', 0.035), ('Neighborhood_Crawfor', 0.028), ('Condition1_Norm', 0.028), ('GarageArea', 0.024), ('TotRmsAbvGrd', 0.023), ('SaleCondition_Normal', 0.023), ('Fireplaces', 0.02), ('BsmtFinSF1', 0.019), ('Neighborhood_StoneBr', 0.019), ('HeatingQC', 0.018), ('Exterior1st_BrkFace', 0.018), ('Neighborhood_NridgHt', 0.017), ('LotFrontage', 0.016), ('GarageCars', 0.015), ('BsmtFullBath', 0.014), ('ScreenPorch', 0.014), ('Exterior1st_MetalSd', 0.014), ('LotArea', 0.012), ('KitchenQual', 0.012), ('BsmtFinType1', 0.009), ('Functional', 0.009), ('WoodDeckSF', 0.009), ('Exterior1st_VinylSd', 0.009), ('Alley', 0.008), ('1stFlrSF', 0.008), ('FullBath', 0.008), ('GarageFinish', 0.008), ('BsmtQual', 0.007), ('HalfBath', 0.007), ('GarageQual', 0.007), ('BedroomAbvGr', 0.006),

Top 5 predictors

('OverallQual', 0.081)
('GrLivArea', 0.073)
('OverallCond', 0.04)
('TotalBsmtSF', 0.035)

('SaleCondition_Partial', 0.035)

No of predictors with 0 coef
count : 150

Observations
- Value of coeffiecients decreased
- Mean of the final predicted sale price decreased.
- More predictor variables become 0, when alpha was 0.001 it was 131 and with alpha 0.002 it is 150

Point 3:

After optimal alphas are doubled

top 5 predictors and their coeff for Ridge are

('OverallQual', 0.067)

('Neighborhood_StoneBr', 0.063)

('Neighborhood_Crawfor', 0.055)

('Exterior1st_BrkFace', 0.045)

('SaleCondition_Normal', 0.044)

Earlier those were

('Neighborhood_StoneBr', 0.097)

('Neighborhood_Crawfor', 0.073)

('OverallQual', 0.066)

('Condition2_Norm', 0.065)

('Exterior1st_BrkFace', 0.06)

We see SaleCondition_Normal as a new predictor in top 5

----

top 5 predictors and their coeff for Lasso are

('OverallQual', 0.081)

('GrLivArea', 0.073)

('OverallCond', 0.04)

('TotalBsmtSF', 0.035)

('SaleCondition_Partial', 0.035)

Earlier those were

('Neighborhood_StoneBr', 0.113)

('GrLivArea', 0.079)

('Neighborhood_Crawfor', 0.079)

('OverallQual', 0.074)

('Neighborhood_NridgHt', 0.056)

We see OverallCond, TotalBsmtSF, SaleCondition_Partial as a new predictors in top 5

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?

ANS:

I will choose lasso model as my final model,

as lasso has reduced coefficient of 131 predictors to zero.

That tells me 131 featues out of 208 are not usedull in prediction.

Whereas Ridge has all features in model parameters.

meaning they are not as useful in predicting the price of the house.

Hence Lasso gives us sparse solution, it also does feature selection for us.

Lasso is also robust to outliers.

As Lasso reduce many feature’s coeff to 0, which tells us which features are not actually usefull in prediction and can be eliminated. Ridge does not do this.

So I will go with the Lasso as my main 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?

ANS:

R2 of train 0.9107066538497152

R2 of test 0.9045344811192915

```
RMSE is:  0.016060296675242187
```

Top 5 predictors

```
('CentralAir', 0.073)
('1stFlrSF', 0.068)
('MSSubClass_SPLIT FOYER', 0.059)
('SaleType_WD', 0.059)
('RoofStyle_Mansard', 0.057)
```

```
  No of predictors with 0 coef count :  126
```

These are the top 5 predictors now

```
('CentralAir', 0.073)
('1stFlrSF', 0.068)
('MSSubClass_SPLIT FOYER', 0.059)
('SaleType_WD', 0.059)
('RoofStyle_Mansard', 0.057)
```

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?

ANS:

Our model should not memorize each and every data point in the train set such that it will mug up the input and output and will not work efficiently with unseen data.

In short model should not overfitt. It should not be too complex.

Also model should not underfit. It should not be too simple.

This can be ensured by many factors like

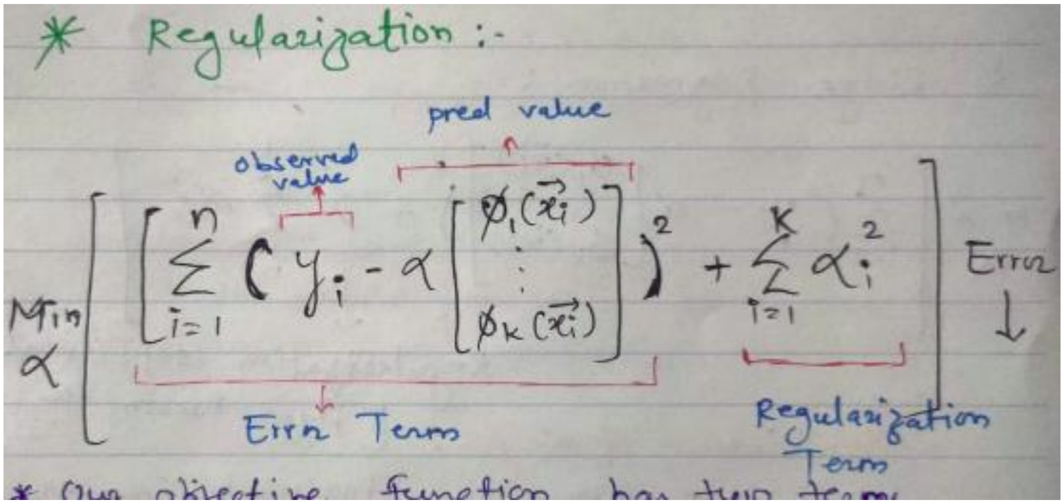
Variance, Bias, R2, Regularization.

We use Regularization techniques to control the model complexity and generalize the model

Model should be able to generalize on the unseen data.

With that in mind, model should not be too complex as well.

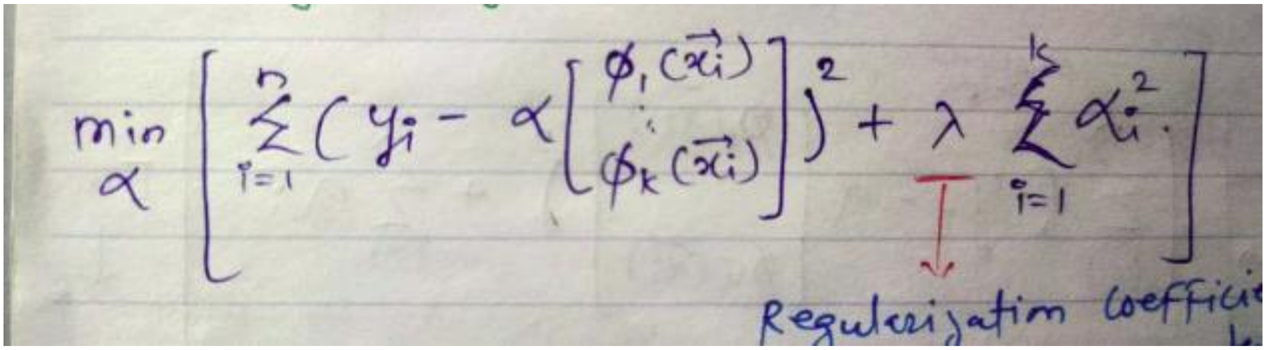
IMAGE 4.1: Regularization



* Regularization :-

$$\min_{\alpha}\left[\underbrace{\left[\sum_{i=1}^{n}\Big(\overset{\text{observed value}}{y_i} - \overset{\text{pred value}}{\alpha\begin{bmatrix}\phi_1(\vec{x_i})\\ \vdots \\ \phi_k(\vec{x_i})\end{bmatrix}}\Big)^2\right.}_{\text{Error Term}} + \underbrace{\sum_{i=1}^{k}\alpha_i^2}_{\text{Regularization Term}}\right]$$

IMAGE 4.2: Regularization Equation

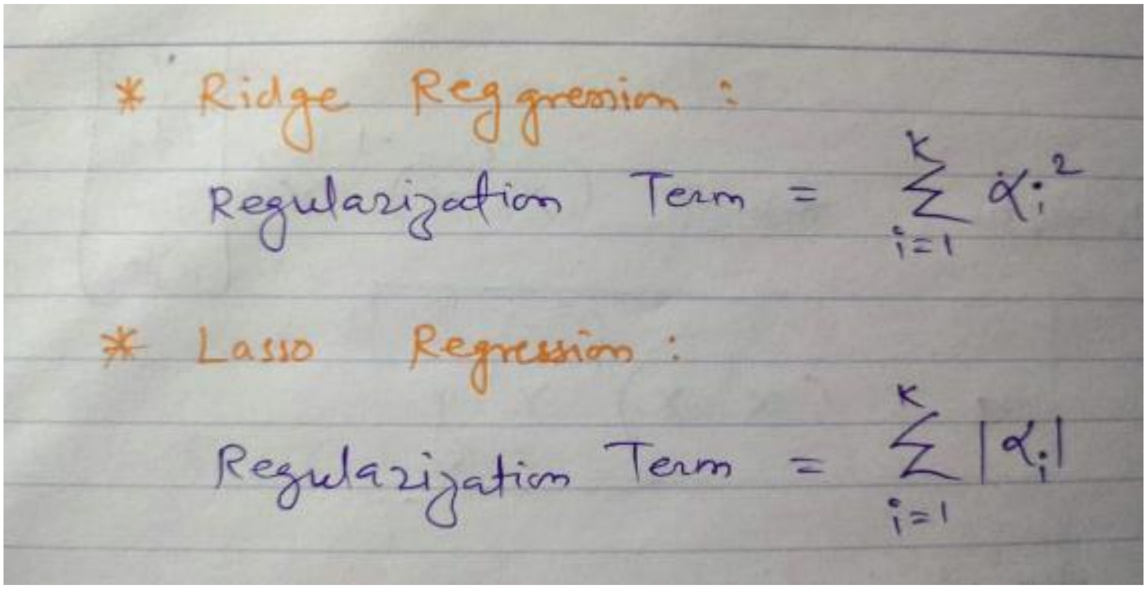


$$\min_{\alpha}\left[\sum_{i=1}^{n}\Big(y_i - \alpha\begin{bmatrix}\phi_1(\vec{x_i})\\ \vdots \\ \phi_k(\vec{x_i})\end{bmatrix}\Big)^2 + \underbrace{\lambda}_{\text{Regularization coeffici}}\sum_{i=1}^{k}\alpha_i^2.\right]$$

Hence if we see the formulas for Ridge and Lasso Regession

these are

IMAGE 4.3: Ridge and Lasso



* Ridge Reggression :

Regularization Term = $\sum_{i=1}^{k} \alpha_i^2$

* Lasso Regression :

Regularization Term = $\sum_{i=1}^{k} |\alpha_i|$

Ridge and Lasso regression apply penalty to the model.

Ridge apply L2 and Lasso apply L1 penalty to the model.

Also we can see alpha the hypermarameter of the modal.

That is there to control the complexity of the modal.

If alpha is too big then we are gicving too much importance to the modal.

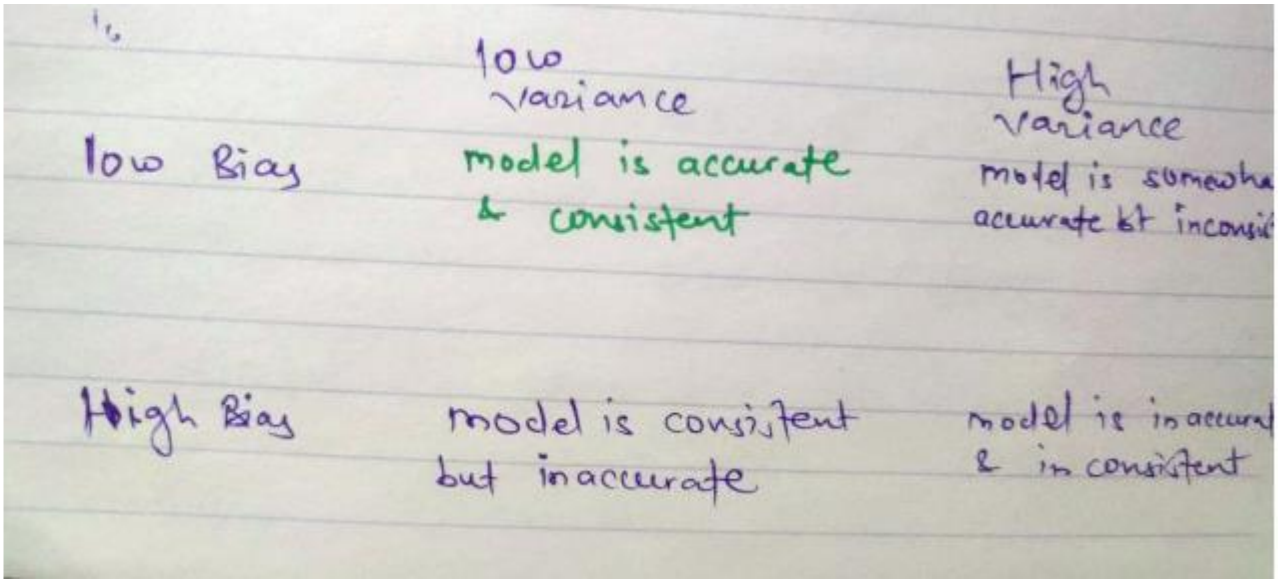
If alpha is too small then the modal is too simple.

Hence we need to find an optimal value of alpha which will generalize the model.

That way we can ensure the modal is generalised.

If modal overfits then the train accuracy will be high but on test data or unseen data it will perform very poorly.

IMAGE 4.4: Bias and Variance



	low variance	High variance
low Bias	model is accurate & consistent	model is somewha accurate bt inconsi
High Bias	model is consistent but inaccurate	model is inaccurat & inconsistent

Variance measures constistency of the modal.

High variance means model is complex and it could overfit.

Bias measure how far are the predicted values from actual values.

Bias measures accuracy of the model.

High Bias means model is too simple and could underfit.

Hence we should aim for Low Bias, Low Variance.