**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:**

# **Ridge Model**

The optimal value of alpha for ridge model seems to be alpha=9. With alpha=9, the test r2 y\_test\_pred matches closely with training r2.

﻿ridge train pred r2\_score: 0.8826345775619088

ridge test pred r2\_score: 0.866495729272287

The top 5 predictor variables in this case are

* ﻿Neighborhood\_NridgHt –
* ﻿Neighborhood\_NoRidge –
* ﻿RoofMatl\_WdShngl –
* ﻿BsmtExposure\_Gd –
* ﻿Neighborhood\_Crawfor –

If I double the alpha to alpha=18, then test r2 does not vary much. Also the top 5 predictor variable remains same.

# **Lasso Model**

For the same alpha=9 or alpha=18 in case lasso the test r2 score y\_test\_pred is lesser than training score. The difference between y\_test\_pred r2 score and y\_train\_pred r2\_score is more than 5 degree.

﻿lasso train pred r2\_score: 0.9343831436100587

lasso test pred r2\_score: 0.8312355842656285

The top 5 predictor variable in this case are

* ﻿RoofMatl\_WdShngl
* ﻿RoofMatl\_Membran
* ﻿RoofMatl\_CompShg
* ﻿RoofMatl\_Metal
* ﻿RoofMatl\_Tar&Grv

This does not change much unless I adjust the alpha for lasso much higher. When I use alpha=72 or beyond then the difference between test and train y prediction become < 5.

﻿lasso train pred r2\_score: 0.8959380007832307

lasso test pred r2\_score: 0.8563338291209158

And with alpha=72, the top 5 predictor variables are

* ﻿RoofMatl\_WdShngl
* ﻿Neighborhood\_NridgHt
* ﻿Neighborhood\_NoRidge
* ﻿RoofMatl\_CompShg
* ﻿Neighborhood\_StoneBr

**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:** Just by looking into the r2\_score and model performance I would choose Lasso as it did automatic feature selection. With alpha=~ 80 the top predictor variables are also explaining overall training set.

**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:** With initial alpha selection alpha=9 or after doubling it to 18 in Lasso the “neighborhood” related variables were not included in top 5 predictor variables. Where as in case Ridge they were there. In housing pricing the area, neighborhood matters a lot for price go up in general.

After adjusting the alpha with a higher alpha for Lasso, then the neighborhood related predictor variables started showing in the top 5 predictor variables.

**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?

Doubt: I didn’t get this question quite, I had to find the model accuracy here ? with metrics.accuracy() sklearn library ?