**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 ridge regression: 10

Optimal valued of lasso regression: 0.001

Ridge regression observation:

Ridge regression with alpha value 10:

R2 value of train data: 0.871473053608503
R2 value of test data: 0.852868745368521

25 feature coefficients for ridge regression as mentioned in table:

Alpha value: 10

| | Feature | Coefficient |
|---|---|---|
| 1 | OverallQual | 0.228 |
| 12 | GrLivArea | 0.187 |
| 5 | BsmtFinSF1 | 0.137 |
| 10 | 2ndFlrSF | 0.129 |
| 8 | TotalBsmtSF | 0.126 |
| 3 | ExterQual | 0.120 |
| 24 | SaleType_New | 0.115 |
| 9 | 1stFlrSF | 0.110 |
| 4 | BsmtQual | 0.106 |
| 13 | GarageArea | 0.102 |
| 15 | Neighborhood_NridgHt | 0.088 |
| 2 | OverallCond | 0.085 |
| 14 | Neighborhood_NoRidge | 0.077 |
| 19 | Exterior2nd_CmentBd | 0.076 |
| 17 | RoofMatl_WdShngl | 0.072 |
| 23 | Heating_Wall | 0.023 |
| 6 | BsmtFinSF2 | 0.021 |
| 0 | constant | 0.000 |
| 21 | Heating_Grav | -0.002 |
| 20 | Heating_GasW | -0.006 |
| 11 | LowQualFinSF | -0.020 |
| 25 | SaleCondition_Partial | -0.021 |
| 22 | Heating_OthW | -0.025 |
| 7 | BsmtUnfSF | -0.028 |
| 18 | Exterior1st_CemntBd | -0.074 |
| 16 | Condition2_PosN | -0.123 |

Ridge regression with alpha value 20:

R2 value of train data: 0.871307552559561
R2 value of test data: 0.8543328475466883

25 feature coefficients for ridge regression as mentioned in table:

Alpha value: 20

| | Feature | Coefficient |
|---|---|---|
| 1 | OverallQual | 0.226 |
| 12 | GrLivArea | 0.187 |
| 5 | BsmtFinSF1 | 0.136 |
| 10 | 2ndFlrSF | 0.128 |
| 8 | TotalBsmtSF | 0.125 |
| 3 | ExterQual | 0.121 |
| 9 | 1stFlrSF | 0.110 |
| 4 | BsmtQual | 0.106 |
| 13 | GarageArea | 0.103 |
| 24 | SaleType_New | 0.094 |
| 15 | Neighborhood_NridgHt | 0.088 |
| 2 | OverallCond | 0.083 |
| 14 | Neighborhood_NoRidge | 0.077 |
| 17 | RoofMatl_WdShngl | 0.071 |
| 19 | Exterior2nd_CmentBd | 0.060 |
| 23 | Heating_Wall | 0.023 |
| 6 | BsmtFinSF2 | 0.021 |
| 0 | constant | 0.000 |
| 25 | SaleCondition_Partial | -0.000 |
| 21 | Heating_Grav | -0.002 |
| 20 | Heating_GasW | -0.006 |
| 11 | LowQualFinSF | -0.020 |
| 22 | Heating_OthW | -0.024 |
| 7 | BsmtUnfSF | -0.027 |
| 18 | Exterior1st_CemntBd | -0.057 |
| 16 | Condition2_PosN | -0.121 |

Observation:

yes, there is the change in r2 values. r2 value train data is slightly decreased but test data is slightly increased but overall not major change. no of feature variable has not changed but their co-efficient are changed because the regularization is the sum of the squared of the co-efficient of the features.

Lasso regression observation:

lasso regression with alpha value 0.001:

R2 value of train data: 0.8713707835676202
R2 value of test data: 0.8536053358864661

21 feature coefficients for ridge regression as mentioned in table:

Alpha value: 0.001

| | Feature | Coefficient |
|---|---|---|
| **12** | GrLivArea | 0.339 |
| **1** | OverallQual | 0.231 |
| **5** | BsmtFinSF1 | 0.165 |
| **3** | ExterQual | 0.119 |
| **4** | BsmtQual | 0.105 |
| **13** | GarageArea | 0.102 |
| **8** | TotalBsmtSF | 0.099 |
| **24** | SaleType_New | 0.094 |
| **15** | Neighborhood_NridgHt | 0.088 |
| **2** | OverallCond | 0.084 |
| **14** | Neighborhood_NoRidge | 0.076 |
| **17** | RoofMatl_WdShngl | 0.071 |
| **19** | Exterior2nd_CmentBd | 0.066 |
| **6** | BsmtFinSF2 | 0.031 |
| **23** | Heating_Wall | 0.023 |
| **0** | constant | 0.000 |
| **10** | 2ndFlrSF | 0.000 |
| **9** | 1stFlrSF | 0.000 |
| **7** | BsmtUnfSF | -0.000 |
| **25** | SaleCondition_Partial | -0.000 |
| **21** | Heating_Grav | -0.001 |
| **20** | Heating_GasW | -0.005 |
| **22** | Heating_OthW | -0.024 |
| **11** | LowQualFinSF | -0.033 |
| **18** | Exterior1st_CemntBd | -0.064 |
| **16** | Condition2_PosN | -0.123 |

lasso regression with alpha value 0.002:

R2 value of train data: 0.8710904228930105
R2 value of test data: 0.8560525531267705

20 feature coefficients for ridge regression as mentioned in table:

Alpha value: 0.002

| | Feature | Coefficient |
|---|---|---|
| 12 | GrLivArea | 0.338 |
| 1 | OverallQual | 0.232 |
| 5 | BsmtFinSF1 | 0.164 |
| 3 | ExterQual | 0.120 |
| 4 | BsmtQual | 0.105 |
| 13 | GarageArea | 0.102 |
| 8 | TotalBsmtSF | 0.100 |
| 24 | SaleType_New | 0.093 |
| 15 | Neighborhood_NridgHt | 0.087 |
| 2 | OverallCond | 0.083 |
| 14 | Neighborhood_NoRidge | 0.075 |
| 17 | RoofMatl_WdShngl | 0.071 |
| 6 | BsmtFinSF2 | 0.030 |
| 19 | Exterior2nd_CmentBd | 0.026 |
| 23 | Heating_Wall | 0.022 |
| 21 | Heating_Grav | -0.000 |
| 0 | constant | 0.000 |
| 10 | 2ndFlrSF | 0.000 |
| 9 | 1stFlrSF | 0.000 |
| 7 | BsmtUnfSF | -0.000 |
| 25 | SaleCondition_Partial | 0.000 |
| 20 | Heating_GasW | -0.004 |
| 22 | Heating_OthW | -0.023 |
| 18 | Exterior1st_CemntBd | -0.024 |
| 11 | LowQualFinSF | -0.032 |
| 16 | Condition2_PosN | -0.121 |

Observation: yes, there is a change in r2 value of train data is slightly decreased but test data is slightly increased but overall, not major change. Number of variables has also changed but the co-efficient has no change since the regularization has the absolute summation of co-efficient of the features.

## 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:

I will choose the lasso regression with alpha value of 0.001 because it helps in feature selection and we have reduced the no of features to 21. Lasso regression also helps in reducing multi collinearity.

**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 current model, the 5 most important variables are below,

1) GrLivArea
2) OverallQual
3) BsmtFinSF1
4) Condition2_PosN
5) ExterQual

After removing the above 5 varibales in the dataset, below are the 5 most important variables,

1) 2ndFlrSF
2) 1stFlrSF
3) TotalBsmtSF
4) Neighborhood_NridgHt
5) BsmtQual

Train data r2 value: 0.8685895645260896
Test data r2 values: 0.8615086351527038

So, r2 values of train and test data have changed after removing the top 5 feature variables but not major difference.

**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:

We should try to build simple model, but it should not be so simple that it will generalized everything in the data set and it will provide biased prediction. Also, simple models are more robust compare to complex models. Simple model can make errors in train data set, but complex model will perform good on train set because of its mimicking property of train set(overfitting). There are ways to make sure that model is robust and generalized which are below.

1) Bias v/s Variance trade off: simple model has high bias but low variance. Complex model has high variance but low bias. Bias refers the accuracy on test data set whereas variance refers the degree of the change in model when we change the training set. We need to find the middle ground b/w bias and variance so that total error is minimized at an accountable threshold value.
2) Regularization: This process is the well know technique to balance b/w simple and complex models by giving some weightage on the coefficient terms of alpha. We can regularized the model with different techniques like ridge, lasso, elastic net.