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?

- The Optimal value of alpha for ridge = 2 and for lasso = 0.0001. With these alphas ,R2 of the model = 0.83.
- Doubling the alpha values using Ridge and Lasso , the optimal value R2 around 0.82 but there is slight change in co-efficient values. These values are derived in assignment submission (code from jupiter notebook).
- Please find the co-efficients are listed below in table between Normal & Doubled alpha for both Ridge & Lasso.

| **Ridge Co-efficient** | | **Ridge Doubled Alpha Co-Efficient** | |
|---|---|---|---|
| Total_sqr_footage | 0.202244 | Total_sqr_footage | 0.149028 |
| GarageArea | 0.110863 | GarageArea | 0.091803 |
| TotRmsAbvGrd | 0.063161 | TotRmsAbvGrd | 0.068283 |
| OverallCond | 0.046686 | OverallCond | 0.043303 |
| LotArea | 0.044597 | LotArea | 0.038824 |
| Total_porch_sf | 0.033294 | Total_porch_sf | 0.033870 |
| CentralAir_Y | 0.028923 | CentralAir_Y | 0.031832 |
| LotFrontage | 0.02337 | LotFrontage | 0.027526 |
| Neighborhood_StoneBr | 0.020848 | Neighborhood_StoneBr | 0.026581 |
| OpenPorchSF | 0.020776 | OpenPorchSF | 0.022713 |
| MSSubClass_70 | 0.018898 | MSSubClass_70 | 0.022189 |
| Alley_Pave | 0.017279 | Alley_Pave | 0.021672 |
| Neighborhood_Veenker | 0.016795 | Neighborhood_Veenker | 0.020098 |
| BsmtQual_Ex | 0.01671 | BsmtQual_Ex | 0.019949 |
| KitchenQual_Ex | 0.015551 | KitchenQual_Ex | 0.019787 |
| HouseStyle_2.5Unf | 0.014707 | HouseStyle_2.5Unf | 0.018952 |
| MasVnrType_Stone | 0.014389 | MasVnrType_Stone | 0.018388 |
| PavedDrive_P | 0.013578 | PavedDrive_P | 0.017973 |
| RoofMatl_WdShngl | 0.013377 | RoofMatl_WdShngl | 0.017856 |
| PavedDrive_Y | 0.012363 | PavedDrive_Y | 0.016840 |

LASSO

| **Lasso Co-Efficient** | | **Lasso Doubled Alpha Co-Efficient** | |
|---|---|---|---|
| Total_sqr_footage | 0.202244 | Total_sqr_footage | 0.204642 |
| GarageArea | 0.110863 | GarageArea | 0.103822 |
| TotRmsAbvGrd | 0.063161 | TotRmsAbvGrd | 0.064902 |
| OverallCond | 0.046686 | OverallCond | 0.042168 |
| LotArea | 0.044597 | CentralAir_Y | 0.033113 |
| CentralAir_Y | 0.033294 | Total_porch_sf | 0.030659 |
| Total_porch_sf | 0.028923 | LotArea | 0.025909 |
| Neighborhood_StoneBr | 0.023370 | BsmtQual_Ex | 0.018128 |
| Alley_Pave | 0.020848 | Neighborhood_StoneBr | 0.017152 |
| OpenPorchSF | 0.020776 | Alley_Pave | 0.016628 |
| MSSubClass_70 | 0.018898 | OpenPorchSF | 0.016490 |
| LandContour_HLS | 0.017279 | KitchenQual_Ex | 0.016359 |
| KitchenQual_Ex | 0.016795 | LandContour_HLS | 0.014793 |
| BsmtQual_Ex | 0.016710 | MSSubClass_70 | 0.014495 |
| Condition1_Norm | 0.015551 | MasVnrType_Stone | 0.013292 |
| Neighborhood_Veenker | 0.014707 | Condition1_Norm | 0.012674 |
| MasVnrType_Stone | 0.014389 | BsmtCond_TA | 0.011677 |
| PavedDrive_P | 0.013578 | SaleCondition_Partial | 0.011236 |

| | |
|---|---|
| LotFrontage 0.013377<br>PavedDrive_Y 0.012363 | LotConfig_CulDSac 0.008776<br>PavedDrive_Y 0.008685 |

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

Based on the derived facts on optimum lambda value in both regression model.

| Ridge Regression Model values | Lasso Regression Model values |
|---|---|
| Lambda = 2 | Lambda = 0.0001 |
| Mean Squared Error – 0.00183 | Mean Squared Error – 0.00186 |
| R2 value = 0.82 | R2 value = 0.82 |

Based on value statistics are almost same between Ridge & Lasso , however Lasso helps in feature reduction (as coefficients of them are zero).
**So I will choose Lasso Regression model for this prediction assignment for 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

The most important predictor variables in Lasso model (after doubling alpha value).

- Total_sqr_footage - 0.204642
- GarageArea - 0.103822
- TotRmsAbvGrd -0.064902
- OverallCond -0.042168
- CentralAir_Y -0.033113

After removing the top five predictor above, built another Lasso model where R2 for this model = 0.73 and MSE = 0.0028

New Top five predictors are listed below.

- LotFrontage - 0.146535
- Total_porch_sf -0.072445
- HouseStyle_2.5Unf -0.062900
- HouseStyle_2.5Fin - 0.050487
- Neighborhood_Veenker -0.042532

**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?

As per **Occam's Razor**

- A model should be as simple as necessary but not simpler than that.
- When in doubt, choose a simpler model.
- Advantages of simplicity are generalisability, robustness, requirement of a few assumptions and less data required for learning

**Bias-Variance Tradeoff**

- Bias measures how accurately a model can describe the actual task at hand.
- Variance measures how flexible the model is with respect to changes in the training data.
- As complexity increases, bias reduces and variance increases, and we aim to find the optimal point where the total model error is the least.

Regularization

- Regularization helps model perform well with unseen data while identifying necessary underlying patterns in it. By adding a penalty term to the cost function used by OLS.
- Ridge and Lasso regression methods, which both allow some bias to get a significant decrease in variance, thereby pushing the model coefficients towards 0.
- In Lasso, some of these coefficients become 0, thus resulting in model selection and, hence, easier interpretation, particularly when the number of coefficients is very large.
- Ideally, we want to reduce both bias and variance because the expected total error of a model is the sum of the errors in bias and variance, as shown in the figure given below.