**Question :**
**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 alpha for ridge regression = 6
Optimal alpha for lasso regression = 100

When alpha value was doubled for both models, it was observed that training accuracy dropped but difference between training and test accuracy was lesser than before. This shows decrease in variance and increase in bias respectively.

Most important predictor after alpha doubled,

for Ridge = OverallQual
for Lasso = 1stFlrSF

**Question :**
**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 to apply Lasso regression for this as there are many feature variables here and Lasso helped me in feature elimination as well. Also, for lasso lambda was higher i.e. 100, which means it would end up regularizing slightly better and have lesser variance.

**Question :**
**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 :

Top 5 most important predictor variables after the change are :

1. TotRmsAbvGrd
2. LotArea_log
3. Neighborhood_NoRidge
4. FullBath
5. BsmtQual_order

To arrive at this, we removed the following predictors :

1. 1stFlrSF
2. Condition2_PosN
3. 2ndFlrSF
4. OverallQual
5. RoofMatl_WdShngl

and then retrained the model with 5 fold validation.

**Question :**
**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 :

1. **Regularization**
   Ridge and Lasso regression are forms of regularization that help in making the model more robust by preventing overfitting. Overfitting occurs when a model learns the noise in the training data rather than the actual underlying patterns, leading to poor performance on new data.
   Ridge Regression adds a penalty on the squared magnitude of the coefficients (L2 penalty). It shrinks the coefficients but does not set them to zero, thus including all features but reducing their impact.
   Lasso Regression adds a penalty on the absolute value of the coefficients (L1 penalty). It can shrink some coefficients to zero, effectively performing feature selection.

   Implications for Accuracy:

   Bias-Variance Trade-off: Regularization introduces a bias (penalty) which reduces variance. This trade-off can lead to a slightly lower training accuracy but typically improves generalization, leading to better performance on the test data.

   Feature Selection (Lasso): By selecting only the most important features, Lasso can simplify the model, potentially improving its interpretability and reducing overfitting.

2. **Cross-Validation**
   Using cross-validation techniques helps in assessing the model’s performance more reliably. It involves splitting the data into multiple folds and training the model on different subsets while validating on the remaining data.

   K-Fold Cross-Validation: The data is split into K subsets. The model is trained on K-1 folds and validated on the remaining fold. This process is repeated K times, and the performance metrics are averaged.

Implications for Accuracy:

Cross-validation helps in estimating the true performance of the model and ensures that the model does not rely on a specific train-test split, thus improving its robustness and generalizability.

3. **Hyperparameter Tuning**
   Finding the optimal value of alpha through cross-validation ensures that the model is neither too simple nor too complex.

   RidgeCV and LassoCV: These methods automate the process of hyperparameter tuning using cross-validation, ensuring the selection of an optimal alpha that balances bias and variance.

   Implications for Accuracy:

   Proper tuning of alpha ensures that the model maintains a balance between underfitting and overfitting, leading to better generalization.
4. **Model Evaluation Metrics**
   Using appropriate evaluation metrics helps in assessing how well the model generalizes.

   R-Squared and Adjusted R-Squared: These metrics indicate the proportion of variance explained by the model. Adjusted R-Squared accounts for the number of predictors, providing a more accurate measure when comparing models with different numbers of features.

   Implications for Accuracy:
   Comprehensive evaluation using multiple metrics ensures that the model's performance is well-understood and robust across different aspects of prediction accuracy.