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

The optimal value of Alpha for both ridge and lasso are 500. Doubling the value of alpha has following impact on Ridge and Lasso:

Impact on Ridge

R2 value slightly drops on training-set and test-set from .86 and .82 respectively to .83 and .81. There is no impact on the top 5 variables selected which are:

OverallQual GrLivArea BsmtQual_Ex TotRmsAbvGrd GarageCars

Impact on Lasso

The bias and variance in the model reduce as the R2 on training and test set before were .90 and .77. After doubling the alpha it gets changed to .87 and .81 respectively. The columns also gets changed before doubling and after doubling

Before

RoofMatl_CompShg GrLivArea RoofMatl_Tar&Grv RoofMatl_WdShngl OverallQual

After

GrLivArea OverallQual BsmtQual_Ex GarageCars BsmtExposure_Gd

Code snippet

```
ridge = Ridge(alpha=2*alpha_ridge)

ridge.fit(X_train, y_train)

# predict
y_train_pred = ridge.predict(X_train)
print('R2 on train set: ', r2_score(y_true=y_train, y_pred=y_train_pred))

y_test_pred = ridge.predict(X_test)
print('R2 on test set: ', r2_score(y_true=y_test, y_pred=y_test_pred))
topCoeff(ridge)
```

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

Lasso always comes with an advantage that it makes the co-efficient values of variables to zero if the variable is least significant for the dependent variable. This reduces the number of variables used to build the model. However, for me Ridge is giving better result. The Bias and variance are lesser in case of Ridge compared to Lasso as we can see the R2 values in question 1.

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 new predictor variables are:

- 1stFlrSF
- GarageCars
- BsmtQual_Ex
- TotRmsAbvGrd
- BsmtFinType1_GLQ

Code Snippet

```
top_lasso_var = ['RoofMatl_CompShg', 'GrLivArea', 'RoofMatl_Tar&Grv', 'Roof
Matl_WdShngl', 'OverallQual']

X_train = X_train.drop(top_lasso_var, axis=1)

X_test = X_test.drop(top_lasso_var, axis=1)

alpha_lasso_2 = model_cv.best_params_['alpha']
print('alpha for lasso: ', alpha_lasso_2)
lasso2 = Lasso(alpha=alpha_lasso_2)

lasso2.fit(X_train, y_train)

# predict
y_train_pred_lasso2 = lasso2.predict(X_train)
print('R2 on train set: ', r2_score(y_true=y_train, y_pred=y_train_pred_las
so2))

y_test_pred_lasso2 = lasso2.predict(X_test)
print('R2 on test set: ', r2_score(y_true=y_test, y_pred=y_test_pred_lasso2
))

print('Total co-efficients:', len(lasso2.coef_))
```

```
print('Removed co-efficients:', list(lasso2.coef_).count(0))
topCoeff(lasso2)
```

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

A model is Robust when it predicts well on the unknown dataset when it is trained on a particular dataset. The accuracy of a robust model is usually good. This is possible when it is less biased. To make sure that a model is robust the pre-processing steps on the dataset should be done rigorously.

A model is Generalizable when it is as simple as possible without much compromising with the accuracy. Generalizable model usually has less variance. Hence once the model is developed it'd accuracy doesn't change much on unknown data as the model is not overfitted.

Impact on Accuracy

• A less robust model doesn't give good accuracy as it results in underfitting.

• A less generalizable model has high impact on accuracy on unknown or unforeseen data