Assignment-based Subjective Questions

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 to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Answer:

According to my model the optimal value of alpha:

1. For Ridge regression :0.5

2. For Lasso Regression :0.0001

If you double the alpha ie

1. For Ridge regression :0.1

2. For Lasso Regression :0.0002

The result obtained are

Ridge model:

<table>
<tr><td>ALPHA=2.0</td><td>ALPHA=1.0</td></tr>
<tr><td>Model parameters and coefficients:
<table>
<tr><th></th><th>Features</th><th>Coefficient</th><th>Mod</th></tr>
<tr><td>0</td><td>LotFrontage</td><td>10.261566</td><td>10.261566</td></tr>
<tr><td>3</td><td>OverallCond</td><td>0.527282</td><td>0.527282</td></tr>
<tr><td>14</td><td>BsmtUnfSF</td><td>0.406426</td><td>0.406426</td></tr>
<tr><td>12</td><td>BsmtFinType2</td><td>0.355562</td><td>0.355562</td></tr>
<tr><td>2</td><td>OverallQual</td><td>0.338037</td><td>0.338037</td></tr>
<tr><td>11</td><td>BsmtFinSF1</td><td>0.322512</td><td>0.322512</td></tr>
<tr><td>9</td><td>BsmtExposure</td><td>0.316265</td><td>0.316265</td></tr>
<tr><td>33</td><td>GarageFinish</td><td>0.303286</td><td>0.303286</td></tr>
<tr><td>73</td><td>LotConfig_CulDSac</td><td>-0.272453</td><td>0.272453</td></tr>
<tr><td>6</td><td>ExterCond</td><td>0.264079</td><td>0.264079</td></tr>
</table></td><td>Model parameters and coefficients:
<table>
<tr><th></th><th>Features</th><th>Coefficient</th><th>Mod</th></tr>
<tr><td>0</td><td>LotFrontage</td><td>10.261566</td><td>10.261566</td></tr>
<tr><td>3</td><td>OverallCond</td><td>0.527282</td><td>0.527282</td></tr>
<tr><td>14</td><td>BsmtUnfSF</td><td>0.406426</td><td>0.406426</td></tr>
<tr><td>12</td><td>BsmtFinType2</td><td>0.355562</td><td>0.355562</td></tr>
<tr><td>2</td><td>OverallQual</td><td>0.338037</td><td>0.338037</td></tr>
<tr><td>11</td><td>BsmtFinSF1</td><td>0.322512</td><td>0.322512</td></tr>
<tr><td>9</td><td>BsmtExposure</td><td>0.316265</td><td>0.316265</td></tr>
<tr><td>33</td><td>GarageFinish</td><td>0.303286</td><td>0.303286</td></tr>
<tr><td>73</td><td>LotConfig_CulDSac</td><td>-0.272453</td><td>0.272453</td></tr>
<tr><td>6</td><td>ExterCond</td><td>0.264079</td><td>0.264079</td></tr>
</table></td></tr>
<tr><td>Accuracy:
<pre>Ridge regression train r2: 0.9239
Ridge regression test r2: 0.7527</pre></td><td>Accuracy
<pre>Ridge regression train r2: 0.9272
Ridge regression test r2: 0.7454</pre></td></tr>
</table>

We don't see much of a difference in top 10 coefficient as the increase of alpha of only by 1. Though the train accuracy decreased the test accuracy increased. Its shows that the second model is more generalized.

Lasso model:

Lasso model:

ALPHA=0.0001

Model parameters and coefficients

	Feature	Coef	mod
0	LotFrontage	10.174821	10.174821
14	BsmtFullBath	0.822796	0.822796
3	OverallCond	0.568885	0.568885
9	CentralAir	0.492011	0.492011
2	OverallQual	0.484268	0.484268
73	Exterior1st_CBlock	-0.479688	0.479688
33	MSZoning_RH	0.379703	0.379703
35	MSZoning_RM	0.297522	0.297522
36	Street_Pave	0.246404	0.246404
20	GarageQual	0.240831	0.240831

Accuracy:

```
Lasso Regression train r2: 0.9280977658497441
Lasso Regression test r2: 0.7373790476938153
```

ALPHA=0.0002

	Feature	Coef	mod
0	LotFrontage	10.174821	10.174821
14	BsmtFullBath	0.822796	0.822796
3	OverallCond	0.568885	0.568885
9	CentralAir	0.492011	0.492011
2	OverallQual	0.484268	0.484268
73	Exterior1st_CBlock	-0.479688	0.479688
33	MSZoning_RH	0.379703	0.379703
35	MSZoning_RM	0.297522	0.297522
36	Street_Pave	0.246404	0.246404
20	GarageQual	0.240831	0.240831

Accuracy

```
Lasso Regression train r2: 0.924842196868562
Lasso Regression test r2: 0.7423837845239192
```

We don't see much of a difference in top 10 coefficient as the increase of alpha of only by 0.001. Though the train accuracy decreased the test accuracy increased. Its shows that the second model is more generalized. The important predictor is MSSubClass

The details about the model is done in the notebook file.

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:

We have built Ridge regression model using optimal alpha as 1 and Lasso regression model using optimal value of alpha as 0.0001.

I would like to apply the Lasso regression model.

The main reason to choose Lasso over ridge regression is, Lasso can perform the Feature selection and reduces the feature coefficient of features which are insignificant implicitly.

This is useful because our business problem has many features and it is not practically optimizable to build model with many features. And Lasso regression model can predict the SalePrice using less features and its accuracy is also in the acceptable range.

Data Modeling and evaluation - Final model

We will make use of Lasso Regression model because it is using less numbers of variables and giving almost the same accuracy. Its more efficient model than Ridge regression model

```
# We have used lasso for building the modle as we got the optimal value of alpha.
lasso = Lasso(alpha = 0.0001)
lasso.fit(X_train,y_train)
y_train_pred = lasso.predict(X_train)
y_test_pred = lasso.predict(X_test)

print(r2_score(y_true = y_train,y_pred = y_train_pred))
print(r2_score(y_true = y_test,y_pred = y_test_pred))
```

```
0.9280977658497441
0.7373790476938153
```

```
#selecting the top 10 variables
lasso_coef.sort_values(by='mod',ascending=False).head(10)
```

	Feature	Coef	mod
0	LotFrontage	10.174821	10.174821
14	BsmtFullBath	0.822796	0.822796
3	OverallCond	0.568885	0.568885
9	CentralAir	0.492011	0.492011
2	OverallQual	0.484268	0.484268
73	Exterior1st_CBlock	-0.479688	0.479688
33	MSZoning_RH	0.379703	0.379703
35	MSZoning_RM	0.297522	0.297522
36	Street_Pave	0.246404	0.246404
20	GarageQual	0.240831	0.240831

Question 3: After building the model, you realized 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:

We have built a model again after deleting the five most important predictor variables. The five most important predictors are:

1. Lot Area: Lot size in square feet
2. Full Bath: Full bathrooms above grade
3. 1st First floor: First Floor square feet
4. External condition: Evaluates the present condition of the material on the exterior
5. MSZoning_RH : Identifies the general zoning classification of the sale(Residential High Density)

	Feature	Coef
0	LotArea	10.205140
10	FullBath	1.029837
6	1stFlrSF	0.606857
1	ExterCond	0.561906
28	MSZoning_RH	0.512930

Question 4: **How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?**

Answer:

According to Occam's razor the model which is simple will give best results. We use the principle of advanced regression technique to make the model more generalizable.

Robust: A model is said to be robust if it is simple, stable and does not change drastically upon changing the training data.

Generalizable: A model is considered generalizable if it does not overfit the training data and gives accuracy in same level with new train and test data.

Moreover, when the model is robust and generalizable it gets less impacted due to outliers and the difference between test and training score is between the acceptable range.

To ensure that the model is robust and generalizable we must not over for our model, do proper outlier and data imbalance analysis. Using Advanced regression techniques find the optimal alpha value and build the model.

Implications on accuracy: A Robust and generalizable model has less difference between Train and test data accuracy. We might not very high train accuracy because it might be a case of overfitting which doesn't make a model robust and can't be used on different set of data. We can use AIC, BIC, R Squarer and Adjusted r squared to measure the accuracy. If there is low AIC, BIC and high Adjusted r square, it shows that the model can perform well on unseen data and it is robust and general.