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 for alpha for

a. Ridge Regression: 20
b. Lasso Regression: 80

If we double the value of alpha for both ridge and lasso, then model will underfit both training and test score will reduce.

Some of the most important predictors during status quo;

	Feature	Linear	Ridge	Lasso
RoofMatl_WdShngl	RoofMatl_WdShngl	711370.029296	15375.684240	78471.697727
Neighborhood_StoneBr	Neighborhood_StoneBr	39093.227219	17652.487714	43867.662066
Neighborhood_NoRidge	Neighborhood_NoRidge	23673.424542	17207.243139	37480.558313
Neighborhood_NridgHt	Neighborhood_NridgHt	19399.086574	21062.547120	37170.600169
Neighborhood_Crawfor	Neighborhood_Crawfor	16295.736329	14582.333805	26069.767866
SaleType_New	SaleType_New	51105.821279	9621.329759	19460.546652
BsmtExposure_Gd	BsmtExposure_Gd	13412.689545	14008.881146	18676.937443
SaleCondition_Alloca	SaleCondition_Alloca	35356.308518	6265.781650	17100.806245
LotConfig_CulDSac	LotConfig_CulDSac	13359.746749	9969.280006	14000.168740

After doubling of alpha value, some of the most important predictors

	Feature	Linear	Ridge	Lasso
RoofMatl_WdShngl	RoofMatl_WdShngl	711370.029296	8672.293026	5.107357e+04
Neighborhood_StoneBr	Neighborhood_StoneBr	39093.227219	11195.302675	3.925759e+04
Neighborhood_NridgHt	Neighborhood_NridgHt	19399.086574	16028.497891	3.539985e+04
Neighborhood_NoRidge	Neighborhood_NoRidge	23673.424542	11274.229329	3.392818e+04
Neighborhood_Crawfor	Neighborhood_Crawfor	16295.736329	10688.619888	2.609960e+04
BsmtExposure_Gd	BsmtExposure_Gd	13412.689545	11411.697678	1.839321e+04
SaleType_New	SaleType_New	51105.821279	7683.703311	1.689152e+04
LotConfig_CulDSac	LotConfig_CulDSac	13359.746749	7849.985650	1.301596e+04
GarageCars	GarageCars	1345.967519	10727.486912	1.220226e+04

So, for conclusion most of the important features still remains the important ones, even after doubling of alpha values.

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:

Score metric for Ridge regression:

```
r2_train_rid :  0.8838236824923261
r2_test_rid :  0.8629747867313836
rss_train_rid :  771924752608.3142
rss_test_rid :  338150236972.91864
mse_train_rid :  760516997.6436594
mse_test_rid :  775573938.0112813
```

Score metrics for lasso regression:

```
r2_train_las :  0.8987391950696177
r2_test_las :  0.8724021293573693
rss_train_las :  672819757689.7795
rss_test_las :  314885481042.5436
mse_train_las :  662876608.5613592
mse_test_las :  722214406.0608798
```

Lasso presents slightly better performance, in addition to the option of removing multiple features as shown:

	Feature	Linear	Ridge	Lasso
HalfBath	HalfBath	-9.333357e+00	1271.932975	0.0
BsmtFinType2_GLQ	BsmtFinType2_GLQ	-8.149549e+03	-1727.991296	-0.0
Functional_Maj2	Functional_Maj2	3.367736e+03	-593.566631	-0.0
Electrical_Mix	Electrical_Mix	-5.735647e+04	-330.948533	-0.0
Electrical_FuseP	Electrical_FuseP	-6.714971e+03	-713.522505	-0.0
Electrical_FuseF	Electrical_FuseF	4.295681e+03	1914.905375	0.0
HeatingQC_Po	HeatingQC_Po	-1.690297e+04	-1183.304679	-0.0
HeatingQC_Fa	HeatingQC_Fa	1.121988e+03	-1777.749154	-0.0
Heating_Wall	Heating_Wall	2.970120e+04	300.903282	-0.0
Heating_OthW	Heating_OthW	-9.272257e+03	-3126.205062	-0.0
Heating_Grav	Heating_Grav	-2.291806e+02	-522.476760	-0.0
Heating_GasA	Heating_GasA	1.442222e+04	252.034014	0.0
BsmtFinType2_NA	BsmtFinType2_NA	-2.986089e+04	-1793.459659	-0.0
BsmtFinType2_LwQ	BsmtFinType2_LwQ	-1.108089e+04	857.026753	0.0
BsmtFinType1_Rec	BsmtFinType1_Rec	1.299445e+03	-129.075672	-0.0
Exterior2nd_Stone	Exterior2nd_Stone	-7.542579e+02	71.100312	-0.0
BsmtFinType1_NA	BsmtFinType1_NA	1.140099e+04	-2569.439059	-0.0
BsmtCond_Po	BsmtCond_Po	7.234917e+04	490.001616	0.0
BsmtCond_Gd	BsmtCond_Gd	-8.058455e+03	-678.430280	-0.0
BsmtQual_NA	BsmtQual_NA	1.140099e+04	-2569.439059	-0.0
Foundation_Wood	Foundation_Wood	-4.079172e+04	-3118.419422	-0.0
Foundation_Stone	Foundation_Stone	2.241229e+03	-165.735491	0.0
Foundation_Slab	Foundation_Slab	-1.120433e+04	-2695.717123	-0.0
ExterCond_TA	ExterCond_TA	-6.828168e+03	886.153733	0.0
ExterCond_Po	ExterCond_Po	-2.054209e+02	-421.359711	-0.0
ExterCond_Gd	ExterCond_Gd	-7.602842e+03	1147.821196	0.0
ExterQual_Fa	ExterQual_Fa	-1.656707e+04	-1471.821581	-0.0
MasVnrType_Stone	MasVnrType_Stone	8.130450e+03	573.889172	0.0
Functional_Min1	Functional_Min1	4.231788e+03	-3253.980936	-0.0
Functional_Min2	Functional_Min2	3.670089e+03	-621.262283	0.0
Functional_Mod	Functional_Mod	3.997798e+03	1777.971659	0.0
Functional_Sev	Functional_Sev	-3.637979e-12	0.000000	0.0
SaleCondition_Family	SaleCondition_Family	4.934030e+03	-1451.515355	-0.0
SaleCondition_AdjLand	SaleCondition_AdjLand	9.948670e+03	909.537561	0.0
SaleType_WD	SaleType_WD	2.458583e+03	-1191.689067	-0.0
SaleType_Oth	SaleType_Oth	3.077488e+04	1428.749671	0.0
SaleType_ConLw	SaleType_ConLw	-3.723135e+03	-1080.667365	-0.0
SaleType_ConLI	SaleType_ConLI	-1.037767e+04	-1762.236044	-0.0
SaleType_ConLD	SaleType_ConLD	2.211441e+04	36.649735	0.0
SaleType_Con	SaleType_Con	1.474512e+04	245.094610	0.0
PavedDrive_P	PavedDrive_P	2.518494e+02	-352.461802	-0.0
GarageCond_TA	GarageCond_TA	4.125026e+03	330.197375	0.0
GarageCond_Po	GarageCond_Po	8.434504e+03	-1752.602313	-0.0
GarageCond_NA	GarageCond_NA	-1.842400e+04	1733.426105	0.0
GarageCond_Gd	GarageCond_Gd	2.212017e+03	1719.812923	0.0
GarageCond_Fa	GarageCond_Fa	3.652447e+03	-2030.834090	-0.0
GarageQual_Po	GarageQual_Po	-1.281363e+05	-969.833909	-0.0
GarageQual_NA	GarageQual_NA	-1.842400e+04	1733.426105	0.0
GarageQual_Gd	GarageQual_Gd	-1.084393e+05	4157.312591	0.0
GarageFinish_NA	GarageFinish_NA	-1.842400e+04	1733.426105	0.0
GarageType_NA	GarageType_NA	-1.842400e+04	1733.426105	0.0
GarageType_CarPort	GarageType_CarPort	3.022100e+04	107.461954	-0.0
GarageType_BuiltIn	GarageType_BuiltIn	2.628672e+04	2127.094397	0.0
GarageType_Basment	GarageType_Basment	2.869616e+04	530.906094	-0.0
FireplaceQu_Po	FireplaceQu_Po	6.137387e+03	103.662467	0.0
FireplaceQu_NA	FireplaceQu_NA	7.459734e+03	-2525.921958	-0.0
FireplaceQu_Fa	FireplaceQu_Fa	-6.070234e+03	-4270.937184	-0.0
MasVnrType_BrkFace	MasVnrType_BrkFace	4.153246e+03	-2467.253813	-0.0
Exterior2nd_Other	Exterior2nd_Other	-1.449825e+04	-135.494332	-0.0
MSZoning_RH	MSZoning_RH	3.384234e+04	442.305396	0.0
Condition1_PosA	Condition1_PosA	6.512116e+03	371.765998	0.0
BldgType_TwnhsE	BldgType_TwnhsE	-8.773512e+03	-1857.648183	-0.0
BldgType_Duplex	BldgType_Duplex	-4.535448e+03	-4068.284961	-0.0
Condition2_RRNn	Condition2_RRNn	1.401661e+03	500.147661	0.0
Condition2_RRAn	Condition2_RRAn	-8.314742e+03	995.783839	0.0
Condition2_RRAe	Condition2_RRAe	-6.226486e+03	-108.074384	-0.0
Condition2_PosA	Condition2_PosA	-4.685717e-09	0.000000	0.0
Condition2_Norm	Condition2_Norm	-4.356727e+03	10763.551506	0.0
Condition2_Feedr	Condition2_Feedr	-1.639842e+04	-69.017416	-0.0
Condition1_RRNn	Condition1_RRNn	1.066052e+04	-562.148525	-0.0
Condition1_RRNe	Condition1_RRNe	1.213311e+04	328.528372	0.0
Condition1_RRAe	Condition1_RRAe	-4.537710e+03	-2604.826124	-0.0
Condition1_PosN	Condition1_PosN	1.516318e+04	-6074.366944	-0.0
Neighborhood_Veenker	Neighborhood_Veenker	-4.684877e+03	-409.696416	0.0
Exterior2nd_MetalSd	Exterior2nd_MetalSd	4.844896e+03	308.735566	0.0
Neighborhood_Timber	Neighborhood_Timber	-9.310706e+03	-5860.772577	-0.0
Neighborhood_SWISU	Neighborhood_SWISU	-2.446336e+03	-2161.209972	0.0
Neighborhood_NPkVill	Neighborhood_NPkVill	2.180723e+04	-436.565458	0.0
Neighborhood_MeadowV	Neighborhood_MeadowV	-1.013896e+04	323.582388	0.0
Neighborhood_CollgCr	Neighborhood_CollgCr	-9.772171e+03	-4699.136607	0.0
Neighborhood_ClearCr	Neighborhood_ClearCr	-7.019718e+03	373.819990	0.0
Neighborhood_BrDale	Neighborhood_BrDale	-2.538818e+03	-652.211177	0.0
Neighborhood_Blueste	Neighborhood_Blueste	6.215532e-08	0.000000	0.0
LotConfig_FR3	LotConfig_FR3	-1.447564e+04	-1374.680103	-0.0
Utilities_NoSeWa	Utilities_NoSeWa	-5.080307e+04	-3150.272717	-0.0
LandContour_Low	LandContour_Low	-1.315333e+04	1433.873319	0.0
MSZoning_RM	MSZoning_RM	2.791850e+04	-2323.318483	-0.0
HouseStyle_1.5Unf	HouseStyle_1.5Unf	1.014683e+04	1023.594706	0.0
HouseStyle_2.5Fin	HouseStyle_2.5Fin	-1.991941e+04	-1195.218632	-0.0
HouseStyle_SFoyer	HouseStyle_SFoyer	-8.007069e+03	507.267502	-0.0
RoofStyle_Gambrel	RoofStyle_Gambrel	1.107238e+04	-809.975948	-0.0
Exterior2nd_ImStucc	Exterior2nd_ImStucc	1.646350e+03	-6.906876	-0.0
Exterior2nd_HdBoard	Exterior2nd_HdBoard	5.881883e+03	1145.199578	0.0
Exterior2nd_CBlock	Exterior2nd_CBlock	-1.746230e-10	0.000000	0.0

So, since we can get better score with improved performance due to decreased computation requirement. I would go with the selection of Lasso 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:

Currently my 5 most important predictors are:

a. RoofMatl: Roof material
b. Neighborhood: Physical locations within Ames city limits
c. SaleType: Type of sale
d. BsmtExposure: Refers to walkout or garden level walls
e. SaleCondition: Condition of sale

Considering that these five are not available in the incoming data, I'll use the next best five predictors, viz

a. LotConfig: Lot configuration
b. BldgType: Type of dwelling
c. GarageCars: Size of garage in car capacity
d. LandContour: Flatness of the property
e. Exterior1st: Exterior covering on house

Ref:

	Feature	Linear	Ridge	Lasso
RoofMatl_WdShngl	RoofMatl_WdShngl	711370.029296	15375.684240	78471.697727
Neighborhood_StoneBr	Neighborhood_StoneBr	39093.227219	17652.487714	43867.662066
Neighborhood_NoRidge	Neighborhood_NoRidge	23673.424542	17207.243139	37480.558313
Neighborhood_NridgHt	Neighborhood_NridgHt	19399.086574	21062.547120	37170.600169
Neighborhood_Crawfor	Neighborhood_Crawfor	16295.736329	14582.333805	26069.767866
SaleType_New	SaleType_New	51105.821279	9621.329759	19460.546652
BsmtExposure_Gd	BsmtExposure_Gd	13412.689545	14008.881146	18676.937443
SaleCondition_Alloca	SaleCondition_Alloca	35356.308518	6265.781650	17100.806245
LotConfig_CulDSac	LotConfig_CulDSac	13359.746749	9969.280006	14000.168740
RoofMatl_CompShg	RoofMatl_CompShg	655909.563694	-449.915425	13197.968462
BldgType_2fmCon	BldgType_2fmCon	2366.844773	4078.934522	12346.164513
GarageCars	GarageCars	1345.967519	12722.622915	12007.705475
LandContour_Lvl	LandContour_Lvl	5890.355903	8935.575137	11122.784963
Exterior1st_BrkFace	Exterior1st_BrkFace	-3763.109304	8418.978533	10658.857197

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:

Bias – Variance Optimization: The idea of this complete exercise if to find a model which is not overfitting, and thus fail to predict in unseen data. But also, to have a model which is generic enough to have a greater predictive power both during learning and in real time.

So, we need to develop a model with optimum bias and variance. Because if bias is higher, then model underfits, and if variance is higher, the model overfits.